METHODS FOR HANDLING MISSING DATA IN A POPULATION BASED COHORT STUDY

Crichton, Siobhan Laura

Awarding institution:
King's College London

METHODS FOR HANDLING MISSING DATA IN A POPULATION BASED COHORT STUDY

THESIS

Presented for the

DEGREE

OF

DOCTOR OF PHILOSOPHY

by

SIOBHAN LAURA CRICHTON

Division of Health and Social Care Research

King's College London

London

2016

Abstract

Background: Missing data are an unavoidable feature in longitudinal studies and inadequate handling can result in bias. The South London Stroke Register (SLSR) follows up participants at three months and annually after stroke. The majority of data collected are categorical in nature. Typically a third of survivors miss each follow-up and the impact of, and 'best' methods for, dealing with these missing data are not clear. The aim of the thesis is to compare and determine the most appropriate methods for handling non-continuous missing data in the SLSR.

Methods: Exploratory analyses identified predictors of incomplete follow-up and informed a simulation study in which the biases associated with prevalence rates of four indicators of poor outcome were estimated and analysis methods compared across four scenarios. Models making differing assumptions about the missing data assessed the impact of missing data on associations between baseline characteristics and outcomes.

Results: Missing data were strongly associated with disability and activity level after stroke and likely *missing not at random* (MNAR). Estimates of prevalence of poor outcomes from available case analyses were relatively unbiased apart from when a strong MNAR assumption was made and outcomes were strongly associated with dropout, with prevalence underestimated by up to 7% points. Bias was reduced after using multiple imputation (MI) with maximum bias of 5% points. There was no evidence that missing data influenced associations between baseline characteristics and outcome.

Conclusions: Some subgroups of the SLSR are at greater risk of non-participation than others but the resulting bias is likely to be minimal. When summarising population outcomes using rates MI should be used in addition to available case analysis. Future work will seek to further quantify potential biases using routinely collected data from GPs to compare responders and non-responders

Contents

List of Figures

List of Tables

Acknowledgements

I would first like to thank my supervisors for all their help and support throughout the thesis. Prof Janet Peacock and Prof Charles Wolfe who provided advice and supervised the later stage of the thesis and Prof Andy Grieve and Dr Michael Toschke who supervised the early stages and the development of the protocol and study design.

I am extremely grateful for support from all colleagues with whom I have worked closely on the South London Stroke Register and particularly Kitty Mohan and Anita Sheldenkar. I would also like to acknowledge all researchers and patients who have contributed to the register since 1995.

I would like to thank other colleagues in the Division of Health and Social Care Research, particularly Dr Patrick White, Dr Alison Wright and Prof Christopher McKevitt, who have all chaired the divisional PhD writing group over the years, and all past and current group members for providing valuable feedback on many chapters in the thesis. Thanks also to all members of the statistics team for their support and guidance.

Finally I would like to thank my family and friends for their support and particularly Kate Wooldrage for proof-reading the thesis.

Abbreviations

AC . . . . . Available case

AMT . . . . . Abbreviated mental test

BI . . . . . Barthel Index

CC . . . . . Complete case

CI . . . . . Confidence interval

EM . . . . . Expectation maximisation

FAI . . . . . Frenchay Activities Index

GCS . . . . . Glasgow Coma Score

GEE . . . . . Generalised estimating equations

GLM . . . . . Generalised linear model

GLMM . . . . Generalised linear mixed model

HADS . . . . Hospital anxiety and depression scale

HADS-A . . . Hospital anxiety and depression, anxiety subscale

HADS-D . . . Hospital anxiety and depression, depression subscale

IPW . . . . . Inverse probability weighting

IQR . . . . . Interquartile range

LOCF . . . . Last observation carried forward

MAR . . . . . Missing at random

MCAR . . . . Missing completely at random

MCMC . . . Markov Chain Monte Carlo

MI . . . . . Multiple imputation

MIGEE . . . Generalised estimating equation with multiple imputation

MMSE . . . . Mini-mental state exam

MNAR . . . . Missing not at random

OR . . . . . Odds ratio

RRP . . . . . Random Recursive Partitioning

PICH . . . . . Primary intracerebral haemorrhage

SAH . . . . . Subarachnoid haemorrhage

sd . . . . . Standard deviation

se . . . . . Standard error

SF-12 . . . . . 12-Item Short Form Health Survey

SF-36 . . . . . 36-Item Short Form Health Survey

SLSR . . . . . South London Stroke Register

WGEE . . . . Weighted generalised estimating equations

WHO . . . . . World Health Organisation

Chapter 1

Background

1.1 Introduction

Population based cohort studies provide valuable data on the natural history of disease and provide a means of estimating the prevalence of risk factors and disease related outcomes in a population of interest [1]. Missing data are a common and often unavoidable feature of such studies [2]. In any study where the aim is to draw valid inferences about a population from a sample, it is important that the sample is representative of the population, or that any differences are taken into account during analysis. If differences are not accounted for, results may be biased. Particularly in longitudinal studies, it is common that some observations may be more likely to be missing than others. Despite this, complete or available case analyses, in which participants with missing data are excluded, are still commonly applied [3–7].

The South London Stroke Register (SLSR) is a population based cohort study which has been collecting data on people with first ever stroke in a defined area of South London since 1995. Participants are followed up at three months, one year and then annually after stroke until death. Some participants have now been followed up for 20 years providing a rich source of information on outcomes following stroke. At any given follow-up point in the SLSR, only 60-70% of participants who are alive

and eligible, complete the follow-up. It is plausible that those who complete the follow-ups may have different characteristics or outcomes than those who do not. Over 200 papers have been published using data from the SLSR, many focusing on identifying predictors or estimating prevalence of poor outcomes after stroke, but it is not known the extent to which incomplete follow-ups may bias results from such analyses.

The issue of non-participation at follow-up is not unique to the SLSR. Similar patterns of non-response have been reported in other stroke and cardiovascular cohorts [8–16] and many non-stroke cohorts [17–21]. Missing data in such studies need to be carefully handled when some participants are more likely to dropout than others. In a systematic review of dropout in studies of elderly populations, Chatfield et al. included studies with up to 12 years of follow-up and found dropout rates, for reasons other than death, of up to 50% [2]. All 25 studies included in the review experienced some dropout. They also examined characteristics associated with dropout. While there was variation between the studies, older age, cognitive impairment and ill-health were consistently associated with increasing likelihood of dropout. Other studies have also found similar associations with age, ill-health and low socioeconomic status well established as being associated with dropout [10,16,17,22–31] . It is possible to adjust for characteristics such as age, but health status can be more problematic in a longitudinal setting. As participants' health may change over the course of follow-up, it may be that it is their health at the time of follow-up which, in part, causes the follow-up to be missed.

The potential association between current health status and completeness of follow-up was also highlighted by Chang et al., who assessed the effect of dropout in a study of depressive symptoms over 10 years [32]. The highest levels of depression were observed in those who dropped out the soonest, with sharp increases immediately prior to dropout. Therefore it was very likely that those who missed a follow-up

were more depressed than those who remained in the study. When the data were analysed using standard statistical models, the associations between baseline characteristics and depression were underestimated compared to those obtained when models which allow for the non-random dropout.

While it is clear that missing data are common in longitudinal studies and that the inappropriate handling of the missing data can lead to biased estimates and a loss of precision [33], many studies are still published in which the impact of missing data is not considered or accounted for. The number and complexity of methods available for dealing with missing data have increased greatly over the last two decades, stemming from work began in the late 1970's when a number of papers were published which first addressed the problems associated with missing data [34–36]. Prior to then most articles either assumed that all observations in a dataset were equally likely to be missing or missing values were 'completed' using values derived without consideration of what the actual values were likely to be [36, 37]. It was not until 1987, which saw the publication of Little and Rubin's "Statistical Analysis with Missing Data" [38], that the development of missing data methods began to pick up pace. Since then, with the availability of increasingly powerful software and hardware, the number and complexity of 'solutions' to the missing data problem have increased [39]. However, there is no one solution which can be applied to all problems and the best method will vary from study to study [40].

Currently available methods range from relatively simple approaches, including single imputation techniques and inverse probability weighting, to more complex multiple imputation and to model based approaches which allow for the inclusion of all observations without the need for imputation [33]. These methods require that the data are missing at random (MAR), i.e. the probability of having missing data depend only on some factors that has been observed in the study. Selection and pattern mixture models offer more complex alternatives that allow for the possibility

that the data are missing not at random (MNAR), i.e. the probability of missing data that depends on some unobserved factor, but they are less intuitive and are computationally intensive [33, 41]. A full discussion of methods for handling missing data is provided in Chapter 2.

Despite the fact that binary and categorical data are common in medical studies, much of the focus of methodological research in missing data methods has been on continuous data with numerous studies demonstrating the properties of, and comparing, various methods. Less focus has been given to the use of methods for handling non-continuous measures in the presence of complex missing data patterns over extended follow-up periods. In the SLSR, at each follow-up health outcomes are collected using a number of validated scales. Though these scales are ordinal in nature they are often categorised prior to analysis. While this is widely acknowledged to be associated with a loss of statistical power [42–44], it remains common practice. Other information collected at follow-up, such as smoking status or perceived recovery from stroke, are categorical in nature. The focus of the thesis will be on the use of missing data methods for non-continuous response data in the SLSR.

In this thesis missing data in the SLSR are described and methods used to adjust for missing data in longitudinal studies are compared. In the remainder of this chapter, before a formal statement of the aims and objectives of this thesis are provided, standard definitions used when discussing missing data are presented. This is followed by the aims and objectives and an overview of the structure of the thesis.

1.2 Missing data mechanisms and patterns

When discussing missing data it is common to describe the missing data 'patterns' and 'mechanisms'. Missing data patterns provide a means of describing the missingness within a dataset and can distinguish between groups with different levels of missing data. Subjects who drop out of a study but have contributed complete

information up until the point of dropout can be described as having a monotonic missing data pattern. Non-monotonic, or intermittent, missingness describes those who are missing at one or more waves but later return to a study. Examples of potential monotonic and non-monotonic missing data patterns in a study with three waves of data collection are illustrated in Table 1.1.

Table 1.1: Possible missing data patterns in a longitudinal study with 3 waves of data collection.

Monotonic			Non-monotonic		
Wave 1	Wave 2	Wave 3	Wave 1	Wave 2	Wave 3
X	X	X	X		X
X	X			X	X
X				X	
					X

X indicates a completed wave

Missing data mechanisms describe the relationship between the likelihood of having missing data and the observed data. Understanding these mechanisms, and the ways in which they can introduce bias into results, is important to ensure appropriate methods are applied when analysing longitudinal data.

Terminology, used to define missingness mechanisms, was first introduced by Rubin in 1976 [36] and modified, into the format which remains popular in statistical literature today, by Little and Rubin in 1987 [38]. The terms commonly given to the three mechanisms described by Rubin are missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

To formally define these mechanisms, let $Y = y_{i,j}$ denote a complete dataset with i subjects and j variables and Y can be partitioned as $Y = Y_{obs}, Y_{miss}$ where Y_{obs}

contains values observed and Y_{miss} those intended to be observed but were not. Also let $M = m_{i,j}$ be an $i \times j$ matrix of missing data indicators where $m_{i,j} = 1$ if $y_{i,j}$ is missing and $m_{i,j} = 0$ otherwise.

1.2.1 Missing completely at random

Missing data are said to be MCAR if the probability of being missing does not depend on any observed or unobserved factors that are related to the data, Y. Therefore, the missing data are considered a completely random subset of all data and all subjects are equally likely to have a missing value. In other words, the missing data are MCAR if

$$\mathrm{P}(M|Y) = \mathrm{P}(M).$$

MCAR data could occur if, for example, a batch of postal questionnaires were sent out with a missing page to a number of participants in a study. When data are MCAR, ignoring missing data would not introduce bias into the results as complete cases are simply a random subset of the larger sample. However, particularly when carrying out research with human subjects, it is likely that some underlying factor will lead to some subjects being more likely to respond than others and thus unlikely that the missing data are ever MCAR.

1.2.2 Missing at random

When data are MAR, all subjects are not equally likely to be observed but the probability of being observed depends only on some observed factor, i.e.

$$\mathrm{P}(M|Y) = \mathrm{P}(M|Y_{obs}).$$

In a study estimating rates of depression, if for example, females were less likely to disclose whether or not they suffered from depression than males but were also more likely to be depressed, then an overall prevalence rate is likely to be an underestimate. However, provided that the missingness does not depend on whether someone

is depressed of not, then the rate within females and males would be unbiased and if sex is known for all participants the overall rate can be adjusted to take into account differing levels of missingness and the data would be considered MAR.

Although data MAR depend on some other factor, as this is observed it can easily be incorporated into analyses, for example, by inclusion as an explanatory variable in a regression model. Provided adjustments are made to include the known factor on which the missingness depends, MAR data will also provide unbiased parameter estimates. As the missingness mechanism therefore does not need to be modelled, MAR and MCAR missing data are also referred to as *ignorable* missingness.

1.2.3 Missing not at random

The biggest difficulty in analyzing incomplete datasets is when the data are MNAR. MNAR implies that the missing data depend on some unobserved factor, i.e.

$$\mathrm{P}(M|Y) = \mathrm{P}(M|Y_{obs}, Y_{miss})$$

In other words there is some unobserved factor, which may not have been measured at all or which may be a factor which was measured for some but not all subjects, which accounts for differences in the probability of missingness. Failing to account for this can lead to unreliable estimates but, as the factor is unknown or only partially observed, doing so presents a challenge. MNAR data can also be described as *non-ignorable* to reflect the need for analytic methods which explicitly allow for the missing data patterns.

While the theoretical distinctions between these mechanisms are clear, in reality, when presented with a dataset with missing values it is impossible to fully distinguish between them. While the observed data will provide enough information to rule out the possibility of a MCAR mechanism, there is no way to determine whether data are MAR or MNAR [33]. For example, in a longitudinal study where

the outcome of interest is level of disability, it may be that those who have the highest disability levels are most likely to drop out. Greater disability is also associated with older age. When examining the relationship between observed data and missingness, the probability of being missing may increase with age. However, when drop out is dependent on level of disability, independently of age, then age cannot fully account for the drop out process and level of disability would therefore be MNAR.

1.3 Aims and objectives

To test the hypothesis that 'the application of suitable analytical methods to a longitudinal study with high levels of missing data will result in improved estimation when compared to available case analysis', the thesis will seek to answer the following aims and objectives.

The overall aim is to compare and determine the most appropriate methods for handling non-continuous missing data in a cohort study with multiple follow-up assessments

The objectives are:

1. To review the development and application of missing data methods over time.
2. To describe patterns and predictors of missing data in the South London Stroke Register
3. To determine the most appropriate method of handling missing data in a 'real' data set with complex missing data patterns.

Specifically, objective 3 will be achieved by addressing the following sub-objectives:

a To compare the performance of missing data methods when estimating prevalence of poor outcome.

b To compare results of analyses of non-continuous outcomes derived from a continuous measure when imputation techniques are applied before and after transformation.

c To compare the performance of missing data methods when identifying predictors of poor outcome.

1.4 Outline of thesis

In the following chapter statistical methods used to analyse non-continuous longitudinal data and for handling missing data will be formally defined and discussed. A review, carried out to summarise the use over time of a number of missing data methods, defined in Chapter 2, is presented in Chapter 3.

The South London Stroke Register (SLSR) is described in detail in Chapter 4 along with results of exploratory analyses focusing on patterns and predictors of incomplete follow-up. The data from the exploratory analyses were then used to design a simulation study to assess the impact of missing data on estimating the prevalence of poor outcomes, and to compare the performance of various missing data methods in the presence of missing data. The methods of the simulation study are described in Chapter 5 and the results presented in Chapter 6.

In a second study, longitudinal models, which make differing assumptions about the missing data mechanisms, were used to identify predictors of poor outcomes in the SLSR. The methods used in this study are presented in Chapter 5 and the results in Chapter 7. Finally, a discussion and conclusions are presented in Chapter 8.

Chapter 2

Longitudinal data analysis and methods for handling missing data.

In this chapter models used in the analysis of longitudinal data, focusing on non-linear outcomes, are discussed. This is followed by a description of methods commonly used for handling missing data in longitudinal studies.

2.1 Longitudinal data analysis

Longitudinal studies in which data are collected on participants at multiple time points are common. Such studies can be prospective, following participants over time, or retrospective with data from historical time points obtained by, for example, participant recall or by extracting information from records [45]. Prospective studies are more common, as they are less likely to suffer from recall bias and do not rely on potentially poor quality historical records [45, 46].

When repeated measurements are made on the same individual a number of times, these measurements are likely to be correlated. When analysing longitudinal data

it is therefore necessary to allow for this within subject correlation [45, 47].

2.1.1 Generalized linear models

Allowing for correlation in repeated measures is commonly achieved by extending traditional generalized linear models (GLMs). GLMs are themselves extensions of the general linear model for normal responses which takes the form

$$\mathrm{E}\left(Y\right) = \mu = \beta\mathbf{X}$$

where Y is the response, μ the mean response, $\beta = (\beta_0, ..., \beta_p)$ a set of unknown parameters and $\mathbf{X} = (X_1, ...X_p)$ a set of explanatory variables. $\mu = \beta\mathbf{X}$ is referred to as the 'linear predictor'. GLMs allow for response variables from distributions other than the normal to be modelled by relating the response variable to the linear predictor through a link function. That is,

$$\mathrm{E}(Y) = g\left(\mu\right) = \beta\mathbf{X}$$

where g is some link function.

2.1.1.1 Logistic regression models

In the thesis, data analysis will focus on the case of binary and ordinal response variables. For binary response data, the logistic regression model is commonly used and is obtained via the logit link function [48].

$$g\left(\mu\right) = ln\left(\frac{\mu}{1-\mu}\right)$$

It is then possible to obtain the conditional probability of response through the fact that $\mu = P(Y = 1|\mathbf{X})$ giving

$$ln\left(\frac{Pr(Y=1|\mathbf{X})}{Pr(Y=0|\mathbf{X})}\right) = \beta\mathbf{X}$$

2.1.1.2 Proportional odds models

For ordinal responses the logistic regression model can be extended to obtain the proportional odds model [49]. In a proportional odds model, where a response variable has C categories, $C-1$ comparisons are made using logistic regression models to obtain a set of cumulative logits [50]. The proportional odds model compares the probability of having an equal or lower response, Y≤c, to the probability of a larger response, Y>c. This is achieved through specifying a series of logits [50, 51]:

$$g_c(x) = logit[P(Y \leq c|\mathbf{X})] = ln\left(\frac{P(Y \leq c|\mathbf{X})}{P(Y > c|\mathbf{X})}\right) = \alpha_c + \beta\mathbf{X}$$

The equations can equivalently be defined as:

$$g_c(x) = logit[P(Y \leq c|\mathbf{X})] = ln\left(\frac{\phi_0(x) + \phi_1(x) + ... + \phi_c(x)}{\phi_{c+1}(x) + \phi_{c+2}(x) + ... + \phi_C(x)}\right) = \alpha_c + \beta\mathbf{X}$$

where $\phi_c(x) = P(Y = c|x)$. Each logit has its own intercept but the β's are the same. This means that the shape of the response curves are assumed to the same, therefore the model is not equivalent to fitting separate logits for $C-1$ response categories [51]. A likelihood function which is a function of the α_c's and β can be constructed and maximised to obtain estimates of the model parameters and allow the cumulative probability of response in each category to be estimated [51]. The conditional probability of response in each category can then be obtained through subtraction of appropriate cumulative probabilities [48, 50].

The model relies on the assumption that the effect of the covariates is the same across all categories, with the differences between the categories being represented entirely by the intercepts, α_c. The Brant test can be used to test the proportional odds assumption [52, 53]. In the test a series of models are fitted to the data which relax the proportional odds assumption and allow the effect of a covariate to differ across response categories. These models are then compared to the proportional odds model using likelihood ratio tests. For each covariate in the model the null hypothesis that the proportional odds assumption holds is tested, with a statistically

significant p-value suggesting violation of the assumption [52].

In this thesis the proportional odds model was used to handle ordinal outcome data, however alternative models also exist. Along with the proportional odds model, the other two most commonly used are the adjacent category and the continuation ratio models [50]. In the adjacent category model each response category is compared to the next with the logits defined as

$$g_c(x) = ln\left(\frac{P(Y=c|\mathbf{X})}{P(Y=c-1|\mathbf{X}}\right) = \alpha_c + \beta\mathbf{X}.$$

On the other hand the continuation ratio model compares a given category to all lower categories, ie the series of logits are defined as

$$g_c(x) = ln\left(\frac{P(Y=c|\mathbf{X}}{P(Y<c|\mathbf{X}}\right) = \alpha_c + \beta\mathbf{X}.$$

2.1.1.3 Multinomial logistic regression models

The multinomial logistic regression model is suitable for analysing categorical outcome data with more than two nominal response categories [50]. As the outcome data do not have to be ordered, the model is suitable for analysing ordinal response variables where the proportional odds assumption does not hold.

For a response variable with categories, 0,1,...,C, a series of C-1 logit link functions are specified as [50]:

$$\begin{aligned} ln\left(\frac{Pr(Y=1|\mathbf{X})}{Pr(Y=0|\mathbf{X})}\right) &= \beta_{\mathbf{1}}\mathbf{X} \\ ln\left(\frac{Pr(Y=2|\mathbf{X})}{Pr(Y=0|\mathbf{X})}\right) &= \beta_{\mathbf{2}}\mathbf{X} \\ &\vdots \\ ln\left(\frac{Pr(Y=C|\mathbf{X})}{Pr(Y=0|\mathbf{X})}\right) &= \beta_{\mathbf{C}}\mathbf{X} \end{aligned}$$

The probability of response in each category is then defined as [50]:

$$
\begin{aligned}
Pr(Y = 0|\mathbf{X}) &= \frac{1}{1 + e^{g_1(x)} + e^{g_2(x)} + ... + e^{g_C(x)}} \\
Pr(Y = 1|\mathbf{X}) &= \frac{e^{g_1(x)}}{1 + e^{g_1(x)} + e^{g_2(x)} + ... + e^{g_C(x)}} \\
Pr(Y = 2|\mathbf{X}) &= \frac{e^{g_2(x)}}{1 + e^{g_1(x)} + e^{g_2(x)} + ... + e^{g_C(x)}} \\
&\vdots \\
Pr(Y = C|\mathbf{X}) &= \frac{e^{g_C(x)}}{1 + e^{g_1(x)} + e^{g_2(x)} + ... + e^{g_C(x)}}
\end{aligned}
$$

Multinomial models then provide an estimate of the effect of a covariate on the probability of response for each of the categories by comparing to a single reference category [50]. This is in contrast to the proportional odds model where a single regression coefficient is estimated for each covariate in the model.

2.1.2 Extensions of generalized linear models to longitudinal data

Approaches which extend these models to allow for correlation in repeated observations can be divided into three separate classes; marginal, transition and subject-specific models with marginal and subject specific models being the most commonly applied [45].

The interpretations of coefficients from marginal and subject specific models differ [54–58]. In a marginal model the average response is modelled separately from the within subject correlation resulting from repeated measurements on the same subject [45]. The coefficients from a marginal model represent the population average response, and are interpreted in the same way as coefficients from models applied to uncorrelated data, for example, data obtained from a cross-sectional study [45,55]. Three assumptions underpin a marginal model, are described by Diggle (2002) [45].

Letting $i = 1, ..., n$ represent subjects from with data are collected at $j = 1, ...J$ points, then the assumptions are:

1. The marginal mean response, $E(Y_{ij}) = \mu_{ij}$ depends on a set of explanatory variables x_{ij} describing the i^{th} subject at the j^{th} time point where $g(\mu_{ij}) = x_{ij}\beta$ where g is an appropriate link function.

2. The marginal variance depends on the marginal mean with $Var(Y_{ij}) = v(\mu_{ij})\phi$ where v is a known variance function and ϕ is a scale parameter.

3. $Corr(Y_{ij}, Y_{ik}) = p(\mu_{ij}, \mu_{ik}, \alpha)$ with p a known function which describes the correlation between Y_{ij}, Y_{ik} as a function of the marginal means and some additional parameters α.

In a marginal model, if all subjects with the same value of a covariate in the model shared the same probability of having a specific outcome, then the estimates from the model would represent the subjects-specific risk. However, when there is variation in risk among subjects with the same value then the coefficients represent the average risk and therefore the population average risk, and not that of individual subjects [54, 55].

Subject-specific models are usually estimated by incorporating random effects into a GLM [45, 54, 55]. Such models assume variation between subjects, and correlation within subjects, is the result of unobserved natural heterogeneity which can be represented by a probability distribution [45, 59]. The effect of covariates and the within subject correlation are modelled using a single equation.

Coefficients from random effects models represent the effect of a covariate on outcome among all subjects with the same level of the random effect [55]. When subjects with the same value of a covariate have the same risk of a given outcome, the coefficient would be equivalent to the marginal, or population average effect. But where there is heterogeneity, the coefficients from such models represent subject specific

effects [55].

Subject specific, or random effects models, are usually fitted using maximum likelihood estimation or conditional likelihood procedures [55]. A number of approaches to the marginal modelling of binary outcomes, for which full likelihood can be derived, exist [60]. However, the complexity of the associated likelihoods make them difficult to evaluate in practice [45]. Instead Generalised Estimating Equations (GEE), first proposed by Liang and Zeger [61], provide an alternative to maximum likelihood estimation.

The third class of models, transition models, allow for dependence between repeated observations by conditioning the response on responses at previous time points [45]. Marginal, random effects and transition models are described in more detail in the following three sections.

2.1.2.1 Generalised estimating equations

Generalised Estimating Equations (GEE) avoid the need for maximum likelihood estimation. In a GEE working assumptions are made with respect to the correlation through specification of a correlation matrix which assumes the correlation between repeated measures is the same for all subjects [61]. Parameters in a GEE are estimated by solving a score equation. For non-normal data, with $i = 1, .., n$ subjects measured at time points $j = 1, ..., J$, the score equation is given as

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^{'}} (A_i^{1/2} C_i A_i^{1/2})^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

with $\boldsymbol{\mu_i} = E(\boldsymbol{y}_i)$, $\boldsymbol{\beta}$ the regression parameters to be estimated, $A_i = A_i(\boldsymbol{\beta})$ a matrix with marginal variances shown on the diagonal and C_i a matrix containing the marginal correlations representing the relationship between repeated measurements [60, 61]. The exact form of the matrix C_i is often unknown and instead assumptions are made about its structure and it is replaced with a working correlation matrix.

Provided that the model has been correctly specified and the appropriate link function used to relate μ to $\boldsymbol{x}_{\boldsymbol{i,j}}^{\mathsf{T}}\boldsymbol{\beta}$ where $\boldsymbol{x_{i,j}}$ is a vector of covariates, then estimated $\hat{\boldsymbol{\beta}}$ obtained from solving the $S(\boldsymbol{\beta}) = \mathbf{0}$ are asymptotically normal with mean $\boldsymbol{\beta}$ and

$$Var(\hat{\boldsymbol{\beta}}) = I_0^{-1} I_1^{-1} I_0$$

where

$$I_0 = \sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i^{'}}{\partial \boldsymbol{\beta}} (A_i^{1/2} C_i A_i^{1/2})^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^{'}} \right)$$

and

$$I_1 = \sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i^{'}}{\partial \boldsymbol{\beta}} (A_i^{1/2} C_i A_i^{1/2})^{-1} Var(y_i) (A_i^{1/2} C_i A_i^{1/2})^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^{'}} \right)$$

[60, 61]. The working correlation matrix can be specified in a number of ways, though independent, exchangeable, AR(1) and unstructured are the most common choices [39].

When data are MCAR if the working correlation matrix differs from the true correlation structure in the data, estimates of $\boldsymbol{\beta}$ will still be unbiased [39, 60, 61]. When data are MAR or MNAR, incorrectly specifying the working correlation matrix can lead to bias in the estimated $\boldsymbol{\beta}$ [39] .

2.1.2.2 Generalised linear mixed models

Generalised linear mixed effects models (GLMM) are extensions of the GLMs described above which incorporate random effects and provide a means of fitting subject specific models [45,59]. GLMMs can allow for correlation within subjects by the inclusion of a random intercept and/or random slopes. A random intercept allows for the repeated measurements from a subject to be higher or lower overall than the average due to some unobserved factor, while a random slope can allow for the rate of change, or the effect of an individual covariate, to vary randomly between subjects due to some unobserved factor [45].

A simple GLMM with a random intercept only can be obtained by specifying the linear predictor in a GLM as:

$$g(\mu_{i,j}) = \beta_0 + \beta x_{i,j}^{\mathsf{T}} + w_i$$

where i is the ith subject, observed at the j^{th} time point, $x_{i,j}$ is a vector of covariates for fixed effects and $w_i \sim N(0, \sigma_1)$ represents the subject specific error term, i.e. the unobserved heterogeneity, and is assumed to be normally distributed [45].

Further random effects random effects allowing for the effect of a factor, such as time, to vary randomly between subjects using a linear predictor of the form:

$$g(\mu_{i,j}) = \beta_0 + \beta x_{i,j}^{\mathsf{T}} + \gamma_i u_{i,j}^{\mathsf{T}} + w_i$$

where γ_i represents a vector of random effects and $u_{i,j}$ is a vector of covariates for random effects. Like w_i, γ_i are assumed to follow a normal distribution with $\gamma_i \sim N(0, \sigma_i^2)$ [45, 59].

Every subject specific model has an equivalent marginal, or population averaged model [62, 63]. It is possible to derive marginal estimates from random effects models in certain situations. Zeger et all, derived a formula for logistic regression random effects models, which use the variance of the random effect to convert the model coefficients to their equivalent marginal estimates.

$$\beta_M \approx \beta_{RE} \left(\left(\frac{16\sqrt{3}}{15\pi} \right) V + 1 \right)^{-1/2}$$

where β_{RE} are the estimates coefficients from a random effects model and V is the variance of the random effect [58].

Both GEEs and GLMMs can be applied to unbalanced repeated measures data meaning all subjects can be included in analyses, even when some have missing data. GLMMs require the data are at least MAR while GEEs make the stricter

assumption that the data are MCAR [45] or, if MAR that the missingness depends only on covariates, which are included in the model, and not on previously observed values of the outcome [64]. Extensions to GEEs have been proposed to allow for valid inferences to be made under MAR. These, along with other methods suitable for use when missing data are present are discussed further in the following sections.

2.1.2.3 Transition models

Transition models, also referred to as conditional models, simultaneously model the average response and time dependence from repeated observations. Models where the response is discrete are also known as Markov chain models [45]. Transition models condition the response on past observed values of the response with the linear predictor given as

$$\mu_{i,j} = \beta_0 + \beta X_{,i} + \sum_{r=1}^{s} \alpha_r f_r(\mathbf{H}_{ij})$$

where $\mathbf{H}_{ij} = Y_{i1}, ..., Y_{ij-1}$ represents responses at time points before the j^{th} time point and $f_r(\mathbf{H}_{ij})$ is a known function [45]. The simplest model is a first order autoregressive model (AR1) where by the response at time j depends on the response at the previous time point only with $\sum_{r=1}^{s} \alpha_r f_r(\mathbf{H}_{ij}) = \alpha_1 Y_{ij-1}$. In a Markov chain model this results in the probability of transitioning to a specific state, or level of the outcome, at time j depends on the state at the previous time point, but not at earlier points. More complex models condition on earlier prior outcomes and the associations with previous states allowed to vary over time [45].

In general Markov chain models can be easily fitted using standard software. However, they were developed for use with data with measurements at regular, equal intervals. Implementing such models becomes more challenging when data are unbalanced and when data is missing. A further potential limitation of such models is that estimates of β coefficients can be highly sensitive to how the time dependence is modelled [45].

2.2 Methods for handling missing data in longitudinal data

A range of methods have been developed to address the problem of incomplete longitudinal data [39, 65]. Simple methods such as complete or available case analyses simply exclude subjects with missing information, while inverse probability weighting uses data from cases without missing information but gives more weight to those observed but most likely to be missing in an attempt to reduce bias. Other methods for handling missing data can be divided roughly into two categories; those which attempt to complete, or impute, the missing data prior to analysis, and those in which missing data are incorporated into the analytic model. Some of the most common methods are described below along with their advantages and potential shortcomings.

2.2.1 Complete and available case analysis

In a complete case analysis any subject with missing data are excluded, while available case analysis, also known as listwise deletion, is slightly less restrictive excluding only subjects with data missing in the variables under consideration for each calculation performed. The major advantage of these methods is that complete datasets are produced allowing for standard statistical methods to be applied [66]. In either approach, when there are many variables under consideration the cumulative effect of missing data can mean that a substantial proportion of subjects in a study are excluded, reducing the power of the methods applied. Further, by excluding some participants biased estimates may result depending on the underlying missingness mechanism [33]. It is possible to achieve unbiased estimates if the data are MAR, with the missingness dependent on a covariate, and appropriate methods are used which make use of the data in the variable on which missingness depends. For example, if dropout from a study was dependent on age alone, adjustment for age in regression analyses would be valid [67]. However, when the missingness is driven by

the outcome, which itself is fully observed and the missingness is only in covariates included in any analyses, results from complete case analysis may be biased [67,68]. Under these conditions methods such as inverse probability weighting (Section 2.2.2) and multiple imputation (Section 2.2.4) can produce valid estimates [67]. Where the missingness is driven by the outcome, which itself is not fully observed then data would be considered MNAR and biases may result when those completing the study are not representative of the complete sample [66].

In a study of cognitive function in over 75 year olds by Dufouil et al, the effect of dropout and death on parameter estimates obtained from longitudinal studies was highlighted [69]. Subjects were assessed using the Mini Mental State Exam (MMSE), which measures cognitive impairment on a scale from 0(severe) to 30(normal), at baseline and in three further waves of data collection. The average score of the non-completers at the last wave at which they were observed was around 3 MMSE points lower than those completing the study. The rate of decline in the non-completers was also accelerated compared to the completers and so, in cases such as this where missing data are associated with poorer outcome, considering only complete cases would suggest higher levels of cognitive function and slower rates of decline than what is actually the case.

A further consequence of complete case analyses is that, even when unbiased estimates can be obtained, there will be a loss of power and precision in the estimation due to the decreasing sample size.

2.2.2 Inverse probability weighting

Inverse probability weighting (IPW) is commonly used in the design and analysis of surveys in which the survey sample contains an over or under representation of subjects for some sub-populations by design [54]. In such designs the probability of being included is known and so analyses can be weighted to produce estimates that

are representative of the target population.

The same methodology can be applied to the missing data problem by using the observed data to estimate the probability of response for each subject and then weighting observed data by the inverse of these probabilities [70]. This results in subjects with a low probability of response being given more weight to adjust for the low response rate from subjects with similar characteristics.

The probability of response is normally estimated using a logit or probit regression model. The dependent variable in the model is an indicator for response with 0=missing and 1=observed. Seaman et al proposed a set of guidelines to guide the development of the missingness model, consisting of five steps [71].

1. Identification of variables for inclusion in the missingness model: Any variables thought a priori to be predictors of missingness should be included along with any strongly associated with outcome. Of those that are predictive of missingness, any that are independent of both the outcome and all variables in the analysis model may be removed.

2. Examine distribution of variables in the missingness model and where possible transform any continuous variables with long tails and extreme values which are likely to be highly influential in the model

3. Fit the model. If there are too many potential predictors identified in step one, use a variable selection procedure, ensuring to force into the model any variables thought to be strongly related to both missing and outcome

4. Check model fit using standard procedures, such as the Hosmer-Lemeshow test.

5. Compare weights in those with and without missing data. If there are particularly large weights for a small number of cases or zero weights in those

with incomplete data the model fit should be further explored and the model refined.

A major drawback of the IPW method is the requirement of specifying the model for missingness. Parameter estimates can be sensitive to the choice of model and miss-specification can lead to biased results [38]. Further, parameter estimates obtained from IPW are less efficient than those from likelihood based analyses [72].

2.2.2.1 Inverse probability weighting in the longitudinal setting

Inverse probability weights can be applied in a longitudinal setting where there is a monotonic missing data pattern. Weights can be derived by fitting a series of models to appropriate subsets of the data. Let $j = 1, ..., J$ be the waves at which data are collected, with $j = 2$ the first time point, D_i at which a individual, i can drop out. The probability of dropping out at time j can then be defined as [41]

$$P(D_i = j) = \begin{cases} P(D_i = j|D_i \geq j) & j = 2 \\ P(D_i = j|D_i \geq j) \prod_{k=2}^{j-1} [1 - P(D_i = k|D_i \geq k)] & j = 3, ..., n_i \\ \prod_{k=2}^{j-1} [1 - P(D_i = k|D_i \geq k)] & j = n_i + 1 \end{cases}$$

with $P(D_i = j|D_i \geq j)$ estimated using probit or logit models applied to individuals still in the study at time $j-1$. Provided that the missingness is monotone, then the missingness models can included values of the outcome measured at previous time points [41].

Where the missing data are not monotonic, or where baseline predictors of missingness are unknown, the Markov randomised monotone missingness (RMM) model has been described for use in this context [70, 73]. Such models are highly computationally intensive and can be complicated to fit, and are generally only suitable for situations where the number of incomplete covariates is small [70].

2.2.2.2 Doubly robust methods

As described above, IPW relies on the model for missingness to be correctly specified and incorrect specification of the model can lead to bias in parameter estimates [33]. Doubly robust methods extends the ideas underpinning IPW and incorporates an imputation model for the expected value of the missing data [74, 75]. Such methods are robust against misspecification of either the imputation model or the missingness model giving an obvious advantage over IPW alone which relies on the correct specification of a single model.

It has been shown that where both the model for missingness and the imputation model are misspecified doubly robust methods can perform worse than non-doubly robust methods, such as simple IPW [76]. Further, where there are highly influential weights with estimated probability of response is close to one for some individuals, the methods can result in substantial bias [77].

2.2.3 Single imputation

Imputation involves the 'filling in' of missing values using some predefined algorithm. There are many different algorithms used to do so even though many have obvious limitations. All single imputation methods result in the same value being imputed for participants which share the same characteristics, for example, a mean score may be applied to participants in a study who are the same age. Consequently, the correlation between the factors used to impute the data and the variable being imputed will very likely be overestimated [66]. Some of the most commonly reported single imputation techniques are described below.

2.2.3.1 Mean, median and mode imputation:

In mean, median and mode imputation the missing values are filled in using, for example, the mean of all observed data or of some subgroup of the population. Estimates of averages or relative frequencies will only be unbiased if data are MCAR

or MAR when estimates are stratified on the factors on which the missing data depends. Other parameters, such as the variance are likely to be underestimated due to the uncertainty in missing data values not being incorporated [38].

2.2.3.2 Hotdeck imputation

Using a hotdeck approach, missing data values are imputed by adopting values from other subjects with similar characteristics and complete data [78, 79]. The use of the term hotdeck is not well defined and there is no standard procedure for identifying the 'deck' of subjects from which an imputation can be drawn [33, 79]. When the parameter of interest is a time constant parameter (i.e. it does not change across different waves of the study) hot-deck imputation has been shown to produce less biased and more efficient estimates than those obtained from available case analysis when data are MNAR and from regression imputation and last observation carried forward when data are MCAR [80, 81]. When the aim is to assess the effect of time on an outcome, the hotdeck approach has been shown to provide poor estimates of time trends [65, 80].

In some situations random hotdeck methods are applied, in which a value is randomly selected from a pool of subjects with similar characteristics. In other cases the average of the values from the pool of similar subjects are used to impute the missing data. Deterministic hotdeck methods, which do not involve any random sampling, use approaches such as nearest neighbour matching to identify a single complete data subject most like each of the subjects with incomplete data [79].

Whatever method is used, all require a donor or pool of potential donors to be identified. To minimise bias and to avoid a loss of precision, variables on which the subject with missing data is matched to a donor should be associated with both the variable with missing data and the likelihood of being missing [79, 82].

A number of measures of 'closeness' between the subjects with complete and incomplete data are commonly used. Let $x_i = (x_{i1}, ..., x_{ik})$ be a vector containing the values of k covariates for subject i on which the matching is to be performed. The maximum deviation between subject i and subject j, being matched on k variables can be defined as

$$d(i,j) = max_k|x_{ik} - x_{jk}|$$

where the variables $x_1, ...x_k$ have been standardised [79].

Another commonly used measure is Mahalanobis distance which is defined as

$$d(i,j) = (x_i - x_j)^T\widehat{Var}(x_i)^{-1}(x_i - x_j).$$

Including categorical covariates in the these measures can be difficult where this is necessary, using a predicted value obtained from applying models developed on complete data to the whole sample may be more straightforward. For continuous variables the distance measure is then given as

$$d(i,j) = (\widehat{Y}(x_i) - \widehat{Y}(x_j))^2$$

where $\widehat{Y} = x_i^{\mathsf{T}}\widehat{\beta}$ is the predicted mean for subject i obtained from a linear model regressing Y, the value to be imputed, on x using subjects with complete data, where x can be continuous or categorical. For binary data predicted probabilities from logistic regression can be used [79].

Once every $d(i,j)$ has been determined for all subjects with incomplete data, then the pool of potential donor subjects can be identified as all those with $d(i,j) < \delta$, for some pre-specified δ. In a nearest neighbour matching the subject for which $d(i,j)$ is smallest would be selected, or where there is a tie then one subject could be randomly selected. The biggest limitation of each of the methods described above is that they rely on complete data across the set of covariates used to determine closeness.

Hotdeck imputation using random recursive partitioning

An alternative approach to determining the closeness of observations is the use of the random recursive partitioning (RRP) algorithm [83]. The RRP algorithm is based on regression trees and uses Monte Carlo methods to derive a proximity matrix which includes a measure of similarity or proximity for each pair of observations. Rather than represent a distance between two observations the proximity measure is interpreted as "the belief of two observations to be equal in covariates" [83].

Regression trees partition data into strata within which the outcome variable, Y, is homogeneous [83]. This is achieved by regressing Y on a set of covariates, initially in the whole dataset. The data are then divided into two nodes based on values of the covariates which minimise heterogeneity between observations in each node [84]. Where the outcome is continuous the deviance from a linear regression model is often used, while the Gini-index, a measure of equality of values with zero indicating that all values are the same and increasing values representing greater inequality, used for categorical outcomes [83]. Each node is then split into two further nodes in a process that continues until the homogeneity within a node is deemed to be sufficient, or when a further split would reduce the sample size below some predefined minimum [83].

The RRP algorithm makes use of the regression tree algorithm. Letting $x_{i,j} = (x_{i,1}, ..., x_{i,p})$ represent a set of covariates for subject i, then for each subject the outcome used in the regression tree algorithm is a random value drawn form a uniform distribution, ie $Y_i \sim U(0,1)$ [83]. For each pair of subjects, a measure of proximity, $\pi_{i,j}^{(1)}$ is defined with $\pi_{i,j}^{(1)} = 1$ indicating that the pair of observations fall within the same node in the regression tree and $\pi_{i,j}^{(1)} = 0$ otherwise.

This process is then repeated R times, with new draws obtained from the uniform

distribution for each of R regression trees. An overall proximity measure for each pair of observations is then obtained by averaging the proximity measure defined above over R replications. That is $\pi_{i,j} = \frac{1}{R}\sum_{r=1}^{R} \pi_{i,j}^{(r)}$. [83].

The RRP algorithm can be used in a hotdeck nearest neighbour imputation approach using the proximity measure to identify nearest neighbours. The advantage of this approach over other distance measures is that regression trees can incorporate subjects with missing data in covariates [83]. As described above other distance measures require that the covariates used to calculate distances are complete meaning that they cannot be applied in many situations.

2.2.3.3 Regression imputation

In regression imputation appropriate regression models for the outcome of interest are fitted using the observed data and used to predict missing values [85]. Models can be cross-sectional, consisting of data observed at one time point, or extended to include data collected from previous study waves. Regression imputation relies on data being complete in the covariates used in the model and so incorporating longitudinal data is only appropriate where there is a monotone missing data pattern.

In general the imputed values from regression imputation will be more variable than those obtained from mean imputation, however, the overall variability in the imputed dataset may still be too small. This can lead to underestimated standard errors, confidence intervals which are too narrow and increased type I error rate [33]. The incorporation on an error term in the imputation step can be used to ensure variation among imputed data is the same as in the observed data [86]. In regression imputation with an added error term missing data are imputed with the predicted value from the same regression model as used in simple regression imputation plus a value randomly chosen from a $N(0, s^2)$ distribution, where s^2 represents the variance of the residuals from the regression model [85].

In a study comparing imputation methods, regression imputation and regression imputation with an error term were found to have similar levels of bias, while regression imputation resulted in underestimated standard errors, as would be expected [85]. While the addition of the error term did not bias results and retained variability, there was found to be large differences between actual and imputed values (measured using absolute and root mean squared deviation) using simulated data. It should also be noted that this study used data representing longitudinal measurements of health status and regression imputation using baseline covariates, with or without the added error term, resulted in greater bias than using imputation techniques that made use of longitudinal measurements [85].

2.2.3.4 Last observation carried forward

Last observation carried forward (LOCF) is probably the simplest longitudinal imputation method and one which is often utilized in clinical trials although relatively uncommon in observational studies. It assumes that, where data are missing, the value that should have been observed will remain constant since the last point at which it was observed. In longitudinal trials it has been shown that LOCF can produce biased treatment effects and inflated type 1 error rates even when the data are MCAR [87–93]. Despite this it remains popular and is the basis for the primary analysis of many trials [94].

2.2.4 Multiple imputation

2.2.4.1 Overview

One of the biggest issues surrounding the single imputation methods is their inability to capture the uncertainty associated with 'guessing' the missing values. For example, where data are MCAR, the mean of the missing data will be equal to the mean of the observed data. However, it is extremely unlikely that missing observations are exactly equal to the mean, which is an assumption made when using mean

imputation. By imputing mean values the standard deviation of the mean for the missing data will effectively be zero and so any estimates of the standard deviation using all data (observed and imputed) will give deflated results.

Multiple imputation attempts to correct for this by repeating the imputation process multiple times, with each imputation consisting of a value randomly drawn from a distribution of likely values determined from the observed data [95–97]. Initially, multiple copies of the observed data sets are created. The imputation is then carried out separately in each data set by randomly drawing a plausible value obtained by fitting an imputation model using the observed data. The model of interest is then fitted using the data in each of these completed data sets and parameter estimates are averaged across imputed datasets. In doing this the variation between the datasets is also utilized and using the methods of Rubin [95] reliable estimates of the errors can be obtained.

There have been several different figures suggested for the optimum number of imputation datasets to be used [98]. Some authors have suggested that as few as five, or even three, are adequate [95]. Others suggest 20 imputations to be sufficient in many cases [99], or that the number of imputations should equal the percentage of data points that require imputation [98, 100].

After creating m imputed datasets analysis is conducted separately on each dataset. Rubin defined a set of rules which can then be used to combine the parameters from the analysis [95]. Letting Q be the estimate of the parameter of interest and V the variance of that parameter, then the overall parameter estimate is taken to be the mean across all imputations, i.e.

$$\overline{Q} = \frac{1}{m}\sum_{i=1}^{m} Q^{(t)}$$

where $Q^{(t)}$ is the parameter estimates from dataset $t = 1, ..., m$. The variance of $\overline{Q}$ is then given as

$$Var(\overline{Q}) = \overline{V} + (1 + m^{-1}) * V_{between}$$

where

$$\overline{V} = \frac{1}{m} \sum_{i=1}^{m} V^{(t)}$$

with $V^{(t)}$ the estimated variance within dataset $t = 1, ..., m$ and $V_{between}$ is the between imputation variance defined as where

$$V_{within} = \frac{1}{m-1} \sum_{i=1}^{m} (Q^{(t)} - \overline{Q})^2$$

Rubin's rules can be used to combine estimates of means, proportions, regression coefficients, linear predictors and area under receiver operating curves. Other parameters such as odds ratios, hazard ratios, correlations and standard deviations can also be combined but first an appropriate transformation must be performed [68]. In general, parameter estimates and their associated standard errors are first derived using Rubin's rules before appropriate hypothesis tests, often based on Wald test statistics, are conducted [68, 98, 101]. Though not often applied in practice, Rubin's rules have been extended to combine likelihood ratio test statistics [68] and methods proposed for the combining of p-values across imputation datasets [68,102].

When data are MAR or MCAR, multiple imputation will provide unbiased estimates of both the effect of interest and of its associated error, provided that the imputation model is correctly specified [103]. Further, there is no loss of power or efficiency as data from all subjects are included [104].

In the following sections approaches to used to create imputed datasets are described generally, and then issues specific to the imputation of longitudinal, or clustered data, are discussed.

2.2.4.2 Specification of the imputation model

Prior to imputing missing data consideration must be given to the variables that are to be included in any imputation model. When the imputation model is misspecified then results can be biased [97]. It is therefore important that the imputation model is consistent, or congenial, with the substantive model to be used in analysis [68]. When a substantive model is used which cannot be derived from the model used to perform the imputations, estimators obtained from the imputed datasets may be inconsistent [68, 105].

All variables that will be included in the substantive model should also be included in the imputation models, this includes both outcomes and covariates. Imputation models for variables which are the dependent variable in the substantive model also need to take account of the structure of the substantive model, e.g. if interaction terms appear in the final analysis they should also be included in the imputation model. Where the substantive model incorporates clustered or repeated measures data, the imputation model should also incorporate such structures [68, 98, 106]. Multiple imputation for longitudinal data is discussed further in Section 2.2.3.5.

In addition to variables in the substantive model, auxiliary variables can be included in the imputation model, resulting in a model which is more complex than the substantive model [68, 98]. Carpenter and Kenward (2012) describe the properties of variable which should be included [68]. The inclusion of auxiliary variables which predict the chance of being missing and the underlying missing values will reduce bias compared to a complete case analysis (for example via a regression model which includes incomplete covariates). Where the auxiliary variable is predictive of the underlying missing values, but not of the likelihood of being missing, then their inclusion will improve efficiency but not reduce bias. On the other hand where the variable is predictive of only the probability of being missing, and not the actually value of the variables being imputed, their inclusion will neither reduce bias nor

improve efficiency and so can be omitted from the imputation model.

2.2.4.3 Imputation strategies

Theoretical justification for multiple imputation requires that the values which are imputed are independent Bayesian draws from the posterior distribution of the missing data given the observed data [41]. In practice, imputation techniques used often provide an approximation to the posterior distribution [41]. Several different algorithms have been proposed and developed for the imputation stage [66].

The earliest approaches to multiple imputation were often based on the assumption that the missing variables could be defined using a multivariate normal distribution [107]. Where there is a monotone missing data pattern then it is possible to approximate the posterior distribution of the multivariate normal distribution using estimates of the mean and covariance matrix obtained from maximum likelihood estimation [41]. While a multivariate normal distribution may be a reasonable assumption for normally distributed variables, such a method requires that non-normal variables be approximated using a normal distribution. It has been argued that the use of a multivariate normal approach is reasonable for the imputation of binary data [101] but simulation studies have highlighted issues when both binary and ordinal variables are imputed [108–113]. In particular, incorporating binary or ordinal variables in this way may mean that imputed data have to be rounded to the nearest valid value prior to analysis, which can potentially introduce bias [110]. Further, nominal variables can not easily be incorporated [41],

2.2.4.4 Multiple imputation using chained equations (MICE)

Often missing data will occur in more than one variable, and multiple imputation using chained equations (MICE), or fully conditional specification (FCS), provides a routine suitable for imputing missing data and can handle variables of different types. [98].

In MICE, imputations are performed using a series of equations, with one for each variable with missing data. Let $z_1, ..., z_k$ be variables that have been selected for inclusion in the imputation model (these may be outcomes or covariates in the substantive model, or variables selected for inclusion as a result of other properties as described above) some or all of which may have missing data. Then the following steps are used to create imputed datasets:

1. All missing values are filled in using simple random sampling, with replacement, from the observed values.

2. The first variable with missing data, say z_1, is regressed on an appropriate subset of variables which include the values imputed in step 1, using an appropriate model applied to all cases without missing data in z_1. The missing values in z_1 are then imputed with a single draw from the posterior predicted distribution of z_1.

3. For the second variable with missing data, say z_2, the process is repeated with the regression model fitted to all cases without missing data in z_2. In this model, the imputed values for z_1 from the previous step are used in place of those obtained in the simple random sample in step 1.

4. The process is then repeated for all other variables with missing data in turn to complete a cycle, or iteration.

5. The process is then repeated several times, starting from step 2, using the dataset in which all variables have been imputed produced at the end of step 4. Using 10 or 20 iterations helps to ensure that the models are stable and not highly influenced by the random sampling used in step 1. At the end of the desired number of iterations, the first imputed dataset is stored.

6. The process is then repeated to give m imputed datasets [98, 107].

To complete the above steps, it is necessary to define; the types of regression models to be used, the set of variables to be used in the imputations, and the number of imputed datasets to be created. As described above, when non-normal data are to be imputed, the assumption of a multivariate normal distribution can, depending on the sample size, distribution of the data, level of missing data and the method used to round the imputed values, result in substantial biases [108–113], which do not occur when a proportional odds model is used instead to perform the imputations [112, 113]. As MICE uses a separate imputation model for each variable, different variable types can be easily handled [98, 107]. Continuous variables are most commonly imputed based on linear regression models, logistic models used for binary variables, proportional odds models for ordinal, multinomial models for nominal variables and count data can be handled using Poisson regression [98]. One of the potential limitations of the MICE approach is that there may be no known joint distribution for the conditional distributions specified in the imputation equations [98]. This can occur when variables of different types (e.g. one continuous and one ordinal) are both being imputed. Although the impact of this has not been extensively studies, existing evidence suggests that the incompatibility is unlikely to cause serious problems [98, 114].

2.2.4.5 Multiple imputation in a longitudinal setting

Multiple imputation offers a flexible approach that can be used to impute both dependent and independent variables prior to fitting the substantive model. The imputation strategies presented above can be applied longitudinal studies with incomplete outcome data. As discussed previously in Section 2.2.4.2 any imputation model must reflect the structure of the final substantive model, and where models are applied to longitudinal data, imputation models should therefore take into account correlations between measurements from repeated measures taken from individual subjects.

In the context of multiple imputation which draws from a multivariate normal distribution, macros have been developed for multilevel imputation with missing data in either level 1 or level 2 variables using MLwiN [115]. This work has been further developed, along with software (REALCOM-IMPUTE) which handles multiple imputation for clustered data with mixed response types. Rather than assuming a multivariate normal distribution for all variable types, a late multivariate latent normal structure is assumed for non-normal data meaning binary, discrete and categorical data can be appropriately incorporated [116]. Where there is a non-linear relationship between the dependent and an independent variable in the substantive model, REALCOM-IMPUTE cannot impute missing values allowing for the non-linear relationship, which may result in the imputation model not being compatible with the substantive model [116].

When using MICE, random effects models for the imputation of linear variables have been described [114]. For other data types alternative solutions have been proposed including the addition of an indicator variable for group as a fixed effect [98]. However, in the longitudinal setting it is often the case that there will be many individuals with only a small number of measurements making this approach unfeasible. It has also been suggested that multilevel structure can be ignored when the relationship between clustered data is not the focus of analysis, but the implications of doing so have not been fully explored [98]. For longitudinal data, outcomes at a given time point can be imputed using information about the same outcome collected at other time points, resulting in dependence between the individuals measurements. Where individuals have many repeated measurements, which may prohibit the inclusion of data from all other time points to be included in each conditional model, an alternative approach has been proposed which imputes missing data using information from the X time points immediately before and after the one at which data are missing [117].

When analysing a longitudinal dataset with incomplete follow-up, multiple imputation and inverse probability weighting both offer solutions where the mechanism is MAR, but both have strengths and limitations. In multiple imputation it is assumed the distribution of the missing data is known conditional on the observed data, while IPW assumes that the drop-out process can be explained by the observed data. It has been argued that IPW is easier for non-statisticians to understand, and that specifying a correct model for missingness may be easier than specifying an imputation model for multiple imputation [70]. An imputation model must take into account the final substantive model structure, and in some cases the final model may not be known, or the dataset may be used for multiple sets of data analysis which may result in imputations being carried out on multiple occasions. With IPW a single set of weights can be derived which can then be used in any future analysis. However, the major limitation of IPW is that, in order to include information from follow-ups which in many cases may be predictive of drop-out from a study, the missingness must be monotone. In reality, many studies will suffer from intermittent missingness as well as drop out, a situation which can easily be handled using multiple imputation.

2.2.5 Model based approaches

2.2.5.1 Maximum likelihood estimation

Parameter estimates in random effects models can be obtained through maximum likelihood (ML) estimation. When there is no missing data present maximum likelihood estimation can provide a relatively simple means of estimating parameters in the model of interest [33]. The EM (expectation maximisation) algorithm, provides a means of estimated maximum likelihood parameters when it is not possible to directly maximise the likelihood [33]. In the presence of missing data the EM algorithm is an iterative process in which missing values are first filled in using appropriate estimates values. Parameter estimates are then derived using this complete dataset. The missing values are then replaces using new values obtained by

assuming that the parameter estimates in the second step are correct. A new set of parameter estimates are then derived from the likelihood using the newly imputed missing values. This process is repeated until the parameter estimates converge [33]. In a random effects model, maximum likelihood estimation means that the correlation between repeated measurements within a subject are used to adjust parameter estimates made using observed data for the unobserved data at each time point [118].

2.2.5.2 Extensions to generalised estimating equations

Generalised Estimating Equations(GEE) offer an alternative to the need for maximum likelihood estimation for marginal models, but require that missing data are MCAR or that the missingness is dependent on covariates only, an assumption which is often unlikely to be true in reality. Robins et al proposed an extension to GEE in which observations are weighted using inverse probability weights (WGEE) resulting in estimates which are valid under a MAR assumption [119–121] .

Prior to fitting a WGEE weights are derived using the approach described in section 2.2.2.1 For each individual $i = 1, ...n$ measured at time $j = 1, ...J$, the probability of being missing at each time point, v_{ij} is determined using a series of logistic regression models and appropriate multiplication of weights as shown in section 2.2.2.1. Then the score equation for a GEE, described in section 2.1.2, becomes

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \frac{I(D_i = j)}{v_{ij}} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^{'}}(j)(A_i^{1/2} C_i A_i^{1/2})^{-1}(j)(\boldsymbol{y}(j)_i - \boldsymbol{\mu}(j)_i) = \mathbf{0}$$

where $I(D_i = d)$ is an indicator variables coded 1 if individual i was observed at time j and 0 otherwise [60, 119]. The model therefore gives greater weight to subjects observed at points where the population of subjects with similar characteristics are overall unlikely to be observed.

Another alternative is the use of MIGEE in which multiple imputation is applied prior to fitting the GEE [60]. In a MIGEE approach, multiple imputation is carried

out using the methods described in section 2.2.4.5. ensuring that the longitudinal nature of the data is represented by the imputation model, for example via the use of MICE with missing outcomes conditioned on the values of the outcomes at other time points. In each of the imputation sets a GEE is fitted and the results are then combined using Rubin's rules.

Both WGEE and MIGEE can provide unbiased estimates under a MAR assumption [60, 89, 121], but as described in sections 2.2.2 and 2.2.4, IPW methods rely on the correct specification of the missingness model and multiple imputation relies on the correct imputation model being used. MIGEE and WGEE therefore also require the correct specification of the appropriate models. Further, in order to construct weights within longitudinal data WGEE can only be applied in situations where the missing data are monotone, whereas multiple imputation is not restricted by the missing data pattern.

This potential for bias resulting from incorrect model specification was demonstrated by Beunckens et al. [60] in a simulation study to compare the performance of WGEE and MIGEE. They simulated missing data patterns in a binary outcome variable to reflect data MAR due to dropout. The performance of both methods was compared across different sample sizes and the robustness of the methods when the model of response in the WGEE and the imputation model in MIGEE was incorrectly specified. WGEE was found to perform poorly in small sample sizes even when the model of response was correctly specified and across all sample sizes when it was incorrectly specified. In contrast, MIGEE were more robust to model misspecification and provided less biased estimates than WGEE when sample sizes were small. Although these studies have demonstrated that WGEE and MIGEE theoretically improve estimation when data were MAR, they both relay on simulated datasets and, in particular for WGEE, the correct specification of the model of response. In reality, missing data patterns are likely to be more complex with intermittent

missingness likely to be present and many real life studies collect data over many more waves of data collection than were represented in these simulations.

2.2.5.3 Missing not at random models

While the methods described above can be used to reduce bias and increase precision when the data are MCAR or MAR the same does not apply when the data are MNAR. A number of methods have been developed which allow for non-random dropout [122]. These models can be split into three distinct types; pattern mixture, selection and shared parameter models with each model making different assumptions about the underlying missing data mechanism [33, 41, 122]. As the models make assumptions about the missingness mechanism, which are untestable given the observed data, MNAR models are most useful as part of a sensitivity analysis. In particular parameter, estimates when compared to results from a MAR model can give an indication as to whether the conclusions hold under a specific MNAR assumption [41].

Pattern mixture, selection and shared parameter models differ in the way in which the full data density is factorised. Using notation set out in Molenbergs and Kenward (2007) [41], let $i = 1, ..., n$ represent n individuals for which it was intended to collect data on an outcome Y at $j = 1, ..., m_i$ occasions with $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{im_1})^T$ a vector of planned responses for individual i. Further, let $\boldsymbol{R}_i = (R_{i1}, ..., R_{im_1})^T$ be a vector of missing data indicators where $R_{ij} = 1$ if Y_{ij} was observed and 0 otherwise. The density for the full set of data can then be summarised as

$$\mathrm{f}(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi})$$

where X_I is a design matrix for the outcome and $\boldsymbol{\theta}$ is a vector of the corresponding parameter estimates, and W_i and $\boldsymbol{\psi}$ are corresponding matrix and parameters for the missingness mechanism.

In pattern mixture models the density is factorised as:

$$\mathrm{f}(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = \mathrm{f}(\boldsymbol{y}_i | \boldsymbol{r}_i, X_i, \boldsymbol{\theta}) \mathrm{f}(\boldsymbol{r}_i | W_i, \boldsymbol{\psi})$$

with the first factor representing the marginal density of the outcome of interest which is conditional on the missingness pattern. The second factor represents the missingness mechanism which does not depend on the outcome.

Selection models assume a different factorisation with

$$\mathrm{f}(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = \mathrm{f}(\boldsymbol{y}_i | X_i, \boldsymbol{\theta}) \mathrm{f}(\boldsymbol{r}_i | \boldsymbol{y}_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}).$$

Here the first factor represents the marginal density of the outcome of interest while the second again represents the missingness mechanism which is this time conditional on the outcome.

Finally, an alternative factorisation gives rise to shared parameter models where

$$\mathrm{f}(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{b}_i) = \mathrm{f}(\boldsymbol{y}_i | \boldsymbol{r}_i, X_i, \boldsymbol{\theta}, \boldsymbol{b_i}) \mathrm{f}(\boldsymbol{r}_i | W_i, \boldsymbol{\psi}, \boldsymbol{b}_i)$$

with $\boldsymbol{b}_i$ representing a vector of random effects where at least one element is shared between both the first and second (ie the model of interest and the missing data mechanism) factors [41].

In the following sections pattern mixture, selection and shared parameter models are discussed in more detail. Following this a discussion of the use of these models as part of a sensitivity analysis is provided.

2.2.5.4 Pattern mixture models

As shown above, the pattern mixture factorisation splits the full data distribution into a marginal distribution for the missing data process and the conditional distribution of the outcome given the missing data process [123]. In order to fit a

pattern mixture model subjects are stratified into distinct groups according to patterns of missingness and models formulated within each group with the effect of covariates allowed to vary between groups. This leads to models which are over specified and parameters which are non-identifiable. While the effect of a covariate is allowed to vary by time of drop-out, once a drop out has occurred, the effect cannot be estimated. In the pattern mixture framework this is usually handled by specifying identifying restrictions which specify the parameters which cannot be estimated as functions of the corresponding parameters estimated in those with complete data [41].

Pattern mixture models can be fitted by adding additional parameters representing drop out time or pattern, and include interactions between pattern and other parameters, to the GLMM described in section 2.1.2.2 [123,124] . For a single covariate of interest, X_1 with a random intercept and subjects arranged into k patterns or groups, the model would take the form

$$\mu_{i,j} = \beta_0 + \beta \boldsymbol{X}_1 + \sum_{m=1}^{k-1} (\beta_0^m \boldsymbol{D}_m + \beta_1^m \boldsymbol{D}_m \boldsymbol{X}_1)) + w_i$$

where $\boldsymbol{D}_m = \boldsymbol{D}_1, ...\boldsymbol{D}_{k-1}$ are dummy variables representing the K groups. If participants with complete data are in group $K = k$ then β_0 and β_1 represent the intercept and slope for completers. β_0^m and β_1^m then indicate the difference between the intercept and slope in group m compared to group k. Where these are significant, it would suggest that the impact of the covariate on the outcome differs according to missing data pattern.

Pattern mixture models provide estimates of effects within groups, usually defined by time of drop-out. Comparing these estimates in this form is useful for assessing the impact of dropout but when comparing with results from alternative models the fact that pattern mixture models do not lead to estimates of overall effects is one of their disadvantages over other MNAR models [123]. Some authors have however combined these estimates to give an overall estimate [124–126]. To do so the pa-

rameter estimates are weighted using the proportion of the population in each of the groups. So for any parameter in the model, the combined estimate is

$$\hat{\bar{\beta}} = \hat{\beta} + \sum_{m=1}^{k-1} \pi^m \hat{\beta}^m$$

where π^m are estimated as the sample proportion in group $m = 1, ...k-1$.

Calculating a standard error is less straight forward and must account for the uncertainty in the estimation of the population proportions estimated by π^m. Hogan and Laird (1997) derived pooled standard error estimates using the delta method for the case where two groups or patterns were included in the model [125]. Letting $\hat{\beta}^{(c)}$ and $\hat{\beta}^{(d)}$ be the estimated parameter of interest in the completers and drop outs, respectively and $\hat{\pi}^c$ and $\hat{\pi}^c$ be the estimated proportion of completers and dropouts. The pooled variance estimate is then

$$\hat{V}(\hat{\bar{\beta}}) = (\hat{\pi}^{(c)})^2 \hat{V}(\hat{\beta}^{(c)}) + (\hat{\pi}^{(d)})^2 \hat{V}(\hat{\beta}^{(d)}) \frac{\hat{\pi}^{(c)} \hat{\pi}^{(d)}}{N} (\hat{\beta}^{(c)} - \hat{\beta}^{(d)})^2.$$

The above approach for combining parameter estimates assumes that associations between covariates of interest in outcome in those who did not complete the study are the same before and after drop-out. Alternative assumptions can be made through the specification of a set of identifying restrictions which are used to specify the conditional distributions of unobserved outcomes given the values of observed outcomes at previous time points. Multiple imputation methods can then be applied which draw from these conditional distributions and to then determine a pooled parameter estimate with an estimate [41].

2.2.5.5 Selection models

As shown above, the selection model is based on factorising the full data density as [41]

$$\mathrm{f}(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = \mathrm{f}(\boldsymbol{y}_i | X_i, \boldsymbol{\theta}) \mathrm{f}(\boldsymbol{r}_i | \boldsymbol{y}_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}).$$

The second factor represents the relationship between the missingness and the observed data, and allows for the presence or absence of missing data (or the probability of being 'selected', ie contributing data, at each time point) to depend on the subjects measurements, which may themselves be missing or observed [41, 127]. Selection models make the assumption that the relationships between a subjects measurements are the same in subjects who were observed and those unobserved, an assumption that cannot be verified given the observed data only [41].

Using a selection modelling approach, the hazard of selection is most commonly fitted using probit or logistic regression [35]. The outcome of interest is then modelled using an appropriate model, for example a GLMM. The two models can then be modelled simultaneously using maximum likelihood estimation or the Heckman two stage approach where the outcome is linear [33, 35]

The biggest limitation of selection models arises through the need to model missingness jointly with model of interest. When there is significant overlap in the factors which predict both missingness and outcome the method can be unreliable [122]. However, unlike pattern mixture models, selection models directly model the effect of covariates on the outcome making them more intuitive.

2.2.5.6 Shared parameter models

Shared parameter models were also defined by Little (1995) as a random effects dependent selection model [128]. They differ from the above approaches in that, rather than assuming the errors are correlated, a random effect is shared between the two models with the data density factorised as [41]

$$\mathrm{f}(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{b}_i) = \mathrm{f}(\boldsymbol{y}_i | \boldsymbol{r}_i, X_i, \boldsymbol{\theta}, \boldsymbol{b_i}) \mathrm{f}(\boldsymbol{r}_i | W_i, \boldsymbol{\psi}, \boldsymbol{b}_i).$$

Applications of shared parameter models are based on the assumption that given the random effects $\boldsymbol{b}_i$, $\boldsymbol{Y}_i$ and $\boldsymbol{R}_i$ are then conditionally independent [41]. A model for the outcome with appropriate random effects can then be specified along with a model

for the missingness. The missingness model which describes the drop out or missing data process then also incorporates some, or all of the random effects. The random effects which are shared between the two models then represents a latent attribute which is associated with both the outcome of interest and the missingness [41]. Where the estimated variance of the random effect is significantly different from zero, this may suggest that this underlying, unobserved process exists and that therefore the data are MNAR [32]. However, as described in the next section and evidence for or against a MNAR assumption based only on observed data must be treated with caution

2.2.5.7 Sensitivity analysis using MNAR models

The MNAR models each make different, untestable, assumptions regarding the underlying missing data mechanism. The models also rely on the assumption that the chosen measurement model is adequate for modelling the outcome of interest in the complete data. However, when the data are partially missing the assumptions underlying the measurement model cannot be tested [41]. MNAR models are therefore best placed to form part of a sensitivity analysis, to assess the robustness of conclusions drawn from the data under possible departures from a MAR assumption, rather than be applied in the primary analysis of any dataset [41].

While it is possible to distinguish between MCAR and MAR patterns, distinguishing between MNAR and other mechanisms is not possible given only observed data. Where there is evidence that a MNAR model fits the data better than a MAR model, this may not necessarily be a result of a missingness mechanism working in precisely the manner assumed by the MNAR model. Additionally, lack of evidence of a MNAR process based on a single MNAR model cannot rule out a general MNAR mechanism, but can only provide evidence against the specific MNAR mechanism hypothesised by the model used. While MNAR models can in theory be used to construct a test of the hypothesis that the data are MAR (for example, by testing

the significance of the shared random effect variances in a shared parameter model or by comparing a selection model to a simpler model which excludes the drop out process), such tests can be highly sensitive to small changes in the dataset. For example, a selection model used to analyse data from 107 cows over two years showed some evidence of a MNAR process, but the removal of just two cows from the dataset altered the conclusions and resulted in no evidence being found [41, 129]. Molenbergs and Kenward (2007) demonstrated how, for the case of incompletely observed longitudinal data, any MNAR model could be reformulated as a MAR model which proved the same fit to the observed data but that could potentially leads to different conclusions [41]. They also, via the use of simulations, summarise the behaviours of a likelihood ratio test statistic used test for a MNAR mechanism in a selection model, and illustrated its atypical behaviour and particular sensitivity to the presence of any unusual longitudinal profiles in the data [41]. Due to these limitations of MNAR models, they should not be used to conduct primary analyses of data and instead used in a sensitivity analysis designed to reflect plausible missingness mechanisms.

2.3 Summary

In this chapter a range of models applied to longitudinal data were described and methods used for handling incomplete data discussed. Methods described in this chapter are referred to throughout the remainder of the thesis. In Chapter 3 trends in the use of these methods are described. The methods of two studies used to explore the impact of missing data in the South London Stroke Register are presented in Chapter 5 with the results of the first study in Chapter 6 and the second in Chapter 7. In the first study the impact of missing data on estimates of the prevalence of poor outcomes after stroke was explored and analyses conducted using complete cases, available cases, inverse probability weighting, single imputation

methods (mean, median, regression and hot-deck) and multiple imputation implemented using chained equations. In the second study a series of models were fitted to the whole SLSR dataset which make differing assumptions about the missing data mechanism. The models used were GEEs, weighted-GEE, MI-GEE, random effects, shared parameter and pattern mixture models.

Chapter 3

Review: Trends in the publication and application of missing data methods

3.1 Abstract

Background: Missing data are common in clinical studies, particularly those following subjects over time. With increasingly powerful hardware and software the number and complexity of methods for handling missing data have increased dramatically. Many studies still apply methods which can lead to biased and imprecise results if the data are not *missing completely at random*.

Methods: A review was conducted to identify longitudinal studies applying missing data methods and articles describing missing data methods. Articles published up to the end of 2009 were included.

Results: The use of multiple imputation(MI) and mixed models increased rapidly after 2005. However, last observation carried forward (LOCF) was common in trials with the numbers published increasing year on year. LOCF was also increasingly used in observational studies and other single imputation methods used in both study types.

Conclusions: Methods such as MI were increasingly applied to deal with missing longitudinal data but sub-optimal single imputation methods which lead to substantial bias remained popular.

3.2 Introduction

The appropriate handling of missing data is of concern in both longitudinal epidemiological studies and clinical trials with dropout rates of up to 50% not unusual in either [2, 130]. Despite the range of, and advances made in, methods for handling missing data available (see Chapter 2 for a detailed description of methods), researchers are thought to often choose methods due to their familiarity or simplicity [39]. Consequently, methods such as complete case analysis and last observation carried forward continue to be applied in the analysis of longitudinal data despite their limitations.

There are numerous reviews which aim to describe missing data methods in a manner which makes them accessible to applied researchers (see for example, [131–133]). Further, many of the modern missing data methods can now easily be implemented using standard statistical software. For example, the PROC MIXED command to fit linear mixed models in SAS has been included since 1996 [134] while similar commands were added to Stata in 2005 [135], SPSS in 2001 [136] and R via several packages including the lme package first released in 1999 [137] followed by lme4 in 2003 which has been under continuous development since [138,139]. Multiple imputation can also easily be handled, with routines available in SAS since 2001 [140], R since at least 2006 [114, 141], SPSS since 2008 [142] and in Stata since 2004 [143] with improved commands introduced in 2011 [144]. Although possible to fit mixed models and carry out multiple imputation prior to the release dates, these packages simplified the process making them more attractive and accessible.

It is widely acknowledged that the 'best' missing data method will vary from study

to study [40]. And while methods have been repeatedly described in detail with illustrated examples, there is limited literature available to aid applied researchers in deciding on the most appropriate methods for individual studies [131]. Many studies do exist which compare the performance of different methods, though these tend to focus on missing continuous data. However, findings differ from study to study with some concluding complete case analyses lead to less bias than other methods in some situations [145] while others show all methods other than MNAR models to be associated with substantial bias when the data are MNAR [146]. Additionally, the majority use simulated data which may make it difficult for researchers to determine how the findings will impact their own analyses.

Historically there has been limited guidance on how missing data should be handled and reported in trials and epidemiological studies published in peer reviewed journals. The CONSORT guidelines, which provide a checklist for the reporting of clinical trials, require dropout to be detailed but do not provide guidance on handling of missing data in analyses [147]. The ICH (International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use) guidelines, which cover all aspects of conducting clinical trials, are also unable to offer any guidance [148, 149]. In ICH E9 Statistical Principles for Clinical Trials, it is stated that effort should be made to avoid missing values. Where this is unavoidable 'sensible' methods should be used. It is stated that missing values should be avoided as far as possible but 'Unfortunately, no universally applicable methods of handling missing values can be recommended'. In 2008 the Food and Drugs Administration (FDA) in the United States commissioned a report, which was published in 2010, from the National Research Council to provide recommendations for reducing and handling missing data in clinical trials [150, 151]. This report provides more detailed guidance on handling unavoidable missing data in phase III trials. They recommend researchers should first develop a plausible set of assumptions regarding the missing data mechanism, and then run an analysis which would

be valid under this assumption. This should then be followed up with a series of sensitivity analyses which allow for departures from the assumed missing data mechanism, using methods such as pattern mixture modelling where appropriate [150].

In the publication of observational studies, many journals (including the British medical journal and the Lancet) now require studies to follow criteria set out in the STROBE guidelines [152]. STROBE suggest the researcher explain how missing data were addressed but no guidance is given on what methods might be appropriate. Although these guidelines do not suggest which methods should be applied, some authors have published suggested guidelines to aid researchers and ensure methods are appropriately applied and reported. For example, Sterne et al. (2009) provide a set of guidelines for the appropriate use and reporting of multiple imputation, published in the British Medical Journal [97]

A number of reviews of the use and reporting of missing data have set out to determine the extent to which appropriate missing data methods are applied and to determine whether this has changed over time with the advances in missing data methodology [3, 153–156]. In 2012 Karaholios et al reported results of a review of the reporting of missing data in large (¿1000 participants) cohort studies indexed in PubMed and published between 2000 and 2009 [3]. Of the 82 included studies, 45 used complete case analysis only, seven used last observation carried forward, five used multiple imputation and 14 did not describe the method used. Missing data were also poorly reported with only 35 reporting the amount of missing data at each follow-up and only three provided details of the reasons for missing data. STROBE guidelines recommend amount and reasons for missing data at each follow up are fully described, but even after their publication in 2007 only one (of nine) studies fully met all criteria described by STROBE.

A study by Jelicic et al. looked at reporting practices in a sample of longitudi-

nal studies in psychological journals published between 2000 and 2006 [155]. Very few were found to apply any missing data methods. Of 57 articles that reported missing data, 84% used complete or available case analyses, 12% used maximum likelihood approaches, one used multiple imputation and one other used a single imputation method. The studies using maximum likelihood were spread across the whole time period giving no evidence of a change in the use of missing data methods.

Another study, focusing on the handling of missing data in educational research compared articles published in 1999 and 2003 [154]. While the proportion of articles reporting the presence of missing data doubled, there was little difference in the use of missing data methods. Again, the majority applied complete or available case analyses. Only six of the 223 articles published in 2003 used multiple imputation or maximum likelihood methods.

The use of methods in randomised controlled trials (RCTs) with missing outcome data, published one year later in 2004, in the Journal of the American Medical Association, the New England Journal of Medicine, the British Medical Journal and the Lancet was reviewed by Wood et al. [153]. In 37 trials which repeatedly measured the outcome, around half used a complete case analysis. A further 19% used last observation carried forward and 11% used a worst case imputation. Only 14% applied a repeated measures analysis (such as GEE or mixed-models) using all observed data.

The same four journals were examined in a review of the use of multiple imputation in medical research [156]. Here all articles published up until the end of 2008 were considered. A total of 49 RCTs and 50 other studies (the majority of which were retrospective reviews) were identified which used multiple imputation, with a substantial increase observed between 2005 and 2008.

These studies suggest low use of methods other than complete or available case

analysis and last observation carried forward up to 2004, with the use of multiple imputation appearing to have increased since 2005. However, these studies have included articles published over short periods of time and in a limited number of journals, making it difficult to determine the extent to which methodological developments in missing data methods have led to a change in the choice of methods used in routine data analysis.

A review was conducted to summarize the use of methods over time in longitudinal studies since 1987 when Little and Rubin's book "Statistical Analysis with Missing Data" [38] was published. Prior to the publication of this book very little consideration had been given to missing data and it was not until after its publication that the development of methodology began to pick up pace. The methods and results of this review are presented in the remainder of this Chapter.

3.3 Aim

The aim of this study was to investigate whether the advances in missing data methodology have translated into an increase in more advanced missing data methods being used in practice.

3.4 Methods

To achieve this aim a search of all literature published between 1st January 1987 and 31st December 2009 was carried out to identify any articles appearing in peer-reviewed journals relating to either the development of methods for handling missing data or longitudinal studies which routinely applied one or more of the methods described in detail in Chapter 2.

3.4.1 Literature identification

Searches of online databases were carried out to identify relevant literature. Data were extracted from PubMed, Embase/Medline and from Web of Science in January 2010. Using PubMed, Embase and Medline it was only possible to conduct searches of titles, abstracts and keywords. Using Web of Science full texts of articles were searched.

The search terms applied to all three databases are summarised in Table 3.1. Articles, either describing or applying one or more missing data methods, were extracted if they included data from an appropriate source (column 1) OR of a specified type (column 2). They were also required to mention missing data (column 3) AND the use of a method appropriate for handling the missing data (column 4). Other variations of the terms included in the table were also used, where appropriate, to ensure alternative spellings and phrasings were included. For example, LOCF, Last Observation Carried Forward and Last-Observation-Carried-Forward. All of the methods listed in the 4^{th} column are described in full in Chapter 2.

A single search was used to identify papers describing methods and the application of methods. This means that articles describing methods must have been identified as being suitable for data arising from one of the sources, or of a specific type, listed in Table 3.1. While there may be further methods or description of methods with application to other types or sources of data, it is unlikely that methods which had never been described in the context of longitudinal data would be routinely applied to such data. Therefore, a single search was deemed to be sufficient to identify articles relating to longitudinal data with incomplete follow-up and which either described or applied one or more of the missing data methods of interest.

Table 3.1: Overview of search terms used to identify articles reporting use or development of missing data methods

Data Source	Data type	Missing Data	Method
Register	Longitudinal	Missing	Complete case
Registry	Follow up	Incomplete	Available case
Survey	Dynamic	Attrition	LOCF
Observational	Panel	Drop out	Imputation
Trial	Repeated measures	Lost to follow up	Multilevel
Experiment	Prospective	Withdrawal	Mixed effects
Cohort	Growth curves	Retention	Random effects
		Response	GEE
		Noncompliance	Maximum likelihood
		Compliance	Shared parameter model
		Adherence	Selection model
		MNAR	Pattern mixture models
		MAR	EM Algorithm
		MCAR	MCMC
			Bayesian

Note: Additional terms were included for alternative spellings or specification of the broad terms above, e.g. "Follow-up" and "Followup" were included in addition to "Follow up".
Abbreviations: LOCF last observation carried forward, GEE generalised estimation equations, EM expectation maximisation, MCMC markov chain monte carlo, MCAR missing completely at random, MAR missing at random, MNAR missing not at random.

3.4.2 Inclusion criteria

Following identification of articles, they were eligible for inclusion in the study if they met one of two sets of inclusion criteria. The first identifies articles relating to the development of missing data methods while the second focuses on the application of missing data methods to longitudinal studies.

Inclusion criteria: Studies relating to methods for missing data

Articles were included if they provided a description of, or focused on, the develop-

ment or performance of methods for handling missing data suitable for application to longitudinal studies. More specifically, articles of the following nature were included:

- General reviews of missing data methods,
- Reviews of use of missing data methods in specific research areas (e.g. Quality of life studies),
- Proposals of new missing data methods,
- Extensions or validations of existing missing data methods or
- Comparisons of multiple missing data methods (using simulated or empirical data).

Only articles published in the English language were included.

Inclusion criteria: Applications of missing data methods

Articles reporting the use of missing data techniques in the routine analysis of studies following participants over time were included. More specifically, articles were required to meet the following criteria for inclusion:

- Repeated measurements taken on two or more separate occasions
- Human participants
- Missing data reported
- At least one method, other than complete or available case analysis, for handling missing data used.

Studies in which repeated measurements were made during a single data collection session or not on humans were excluded in order to gain an overview of the use of methods in studies in which non-random dropout or loss to follow-up was possible. Again, only articles published in the English language were included.

3.4.3 Data extraction

All identified literature was reviewed to determine eligibility using the above criteria. Initially titles were surveyed and any articles which clearly did not meet inclusion criteria from the title were excluded. For the remaining articles, the abstract was first used to identify those which met the inclusion criteria and, where the abstract provided insufficient information, the full text was consulted.

A database was created in Microsoft ACCESS to aid coding of the articles. All articles were initially defined as 'not eligible', 'method' or 'application'. Where an article was not eligible, the main reason was provided. For 'methods' articles, all missing data methods discussed were recorded. The nature of the article was also recorded as one of the following: general review, review specific to a research area, description of new or extension of existing methods or comparison of methods. In the 'application' group, all missing data methods applied in the study were recorded and the source of the data (i.e. trial or observational study) also noted.

3.5 Analysis

Data were exported into STATA 11ME. The frequency of method and application articles were calculated by year of publication. Histograms were constructed to assess trends in the development and use of each missing data method over time. Comparisons were also made between the use of methods in studies with data arising from clinical trials and those from other observational studies.

Crude estimates of the total number of all observational studies and randomised controlled trials (RCTs) published over the same time period were extracted using an online trend database (http://dan.corlan.net/medline-trend.html). The database provides the total number of articles, by year of publication, indexed in medline and meeting a set of inclusion criteria. To identify RCTs articles (randomi* AND con-

trolled AND trial) was used as a search strategy and ((observational OR cohort OR regist* OR survey) AND (follow up OR longitudinal OR panel OR repeated measures)) to obtain estimates for observational studies. Bar charts were constructed to allow comparisons between publication rates of studies which applied missing data methods and the overall numbers published. These data were extracted in July 2011 and contained all articles published from 1st January 1987 to the 31st December 2009.

3.6 Results

The results of the electronic database searches are summarised in the flow diagram in Figure 3.1. Across all three databases, a total of 11094 articles were identified, of which 8534 were unique. Nine hundred and twelve articles related to the development of missing data methods and 1687 were studies which applied one or more of these methods. Table 3.2 summarizes the main focus of the remaining 5935 articles that were not included. The most common reason was that they were systematic reviews or meta-analyses. Descriptions of statistical methods not directly related to the missing data problem or not applicable to analysis of longitudinal data were common. Meanwhile, case studies and small case series, in which outcomes are summarised using a series of rates and proportions, and studies not on human participants, and so not subject to random dropout, were also frequently excluded.

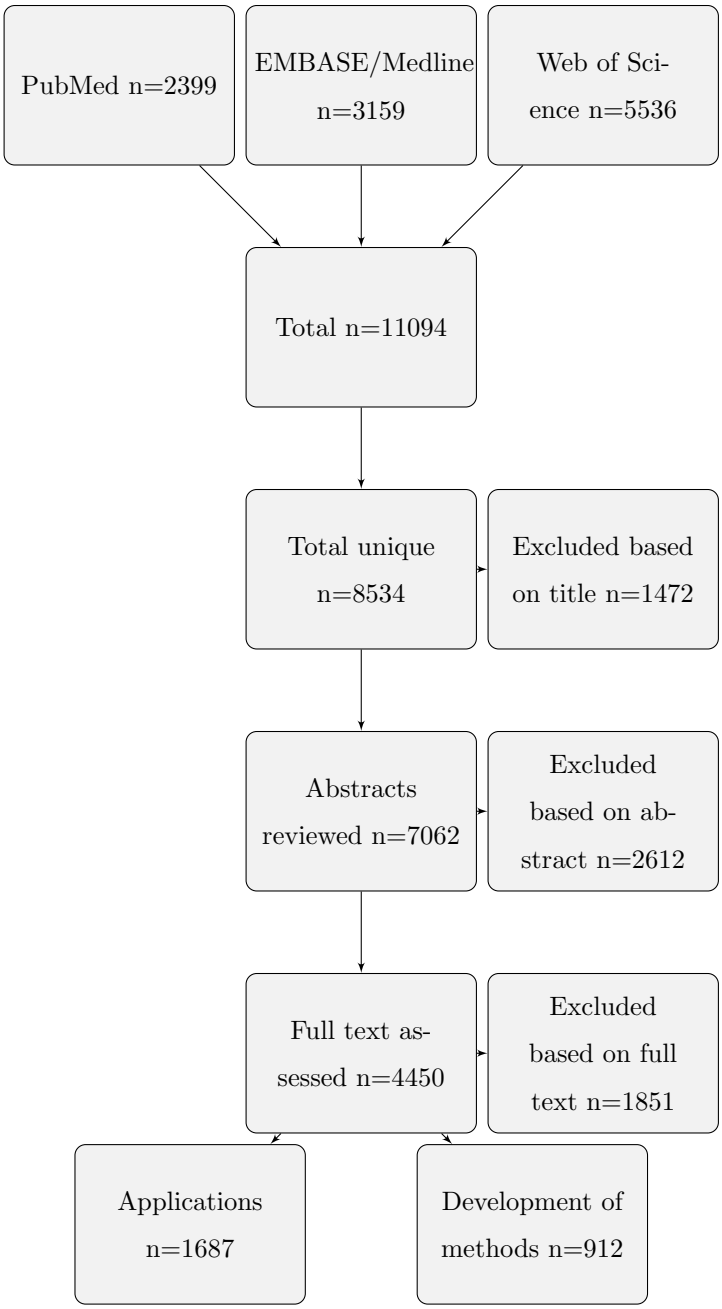


Figure 3.1: Flow diagram of study selection for inclusion in review

Table 3.2: Reasons for exclusion

Reason	N	Reason	N
Methodological		**Not Human**	
Not missing data methods	637	Animal	286
Not applicable to longitudinal data	112	Environmental	80
Sample size and power calculations	25	Other	534
Reviews		**Other**	
Systematic review and meta-analyses	1011	Dynamic Models	60
Other reviews and consensus statements	230	Cost	231
Follow up studies		Cellular	52
Adherence to interventions	26	Audit	22
Administrative databases	100	Time Series	148
Case Series	552	Genetic/Microarray	111
Case Study	182	Sensitivity/Specificity	145
Complete case analysis only	254	Other	398
No missing data reported	107		
Single data collection point	518		
Survival analysis	95		

The total number of application and method papers using or describing various missing data methods are summarised in Table 3.3. The most common method to be both described and used in practice was mixed model analysis which was included in 22.8% of the method articles and used in 45.6% of the applications. Last observation carried forward (LOCF) was the second most commonly applied method, being used in 39.2% of studies. However, it featured in only 6.4% of the method articles and the majority focused on the limitations of the method. Conversely, more advanced and computationally intensive methods, such as likelihood based approaches and MNAR models, featured in 12.9 and 12.5% of the method papers but were applied in only 0.8 and 1.7% of studies, respectively.

Table 3.3: Summary of included articles

Database	Methods	Applications
Total	912	1687
Weighting	23(2.5%)	9(0.5%)
LOCF	58(6.4%)	658(39.2%)
Other Imputation	68(7.5%)	58(3.5%)
Multiple Imputation	110(12.1%)	119(7.1%)
Generalized Estimating Equations	46(5.0%)	24(1.4%)
Mixed models	208(22.8%)	766(45.6%)
Other likelihood based	118(12.9%)	14(0.8%)
MNAR models	114(12.5%)	28(1.7%)
Other	223(24.5%)	37(2.2%)

Abbreviations: LOCF last observation carried forward, MNAR missing not at random.

The proportion of all 912 method articles and 1687 applications published by calendar year are displayed in Figure 3.2. While the number of methodological studies has increased linearly over time since 1987, the number reporting the routine use of these methods has grown exponentially, with the majority of the studies being published after 2000 highlighting the delay between methods being developed and being adopted into routine data analysis.

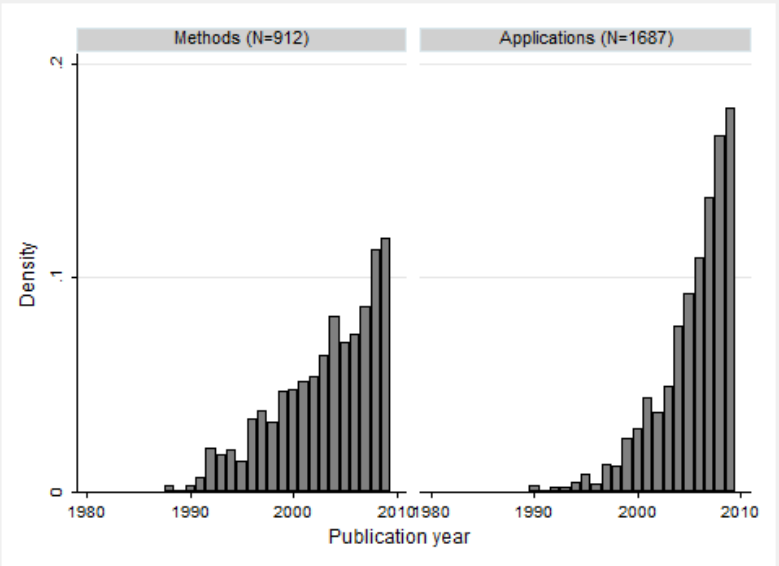


Figure 3.2: Distribution of articles reporting the development and application of missing data methods over time

This trend is further highlighted in Figures 3.3 to 3.8 which illustrate the publication rates by calendar year for the six most commonly applied missing data methods: last observation carried forward, other single imputation methods, multiple imputation, mixed models, generalised estimating equations and MNAR models. For each of the missing data methods, the publication rate of method articles shows a steady increase while the number of applications increases at a more rapid rate, starting several years after the initial increase in the number of method papers. While there has been an increase in modern missing data methods such as multiple imputation and mixed model analysis, the application of last observation carried forward and other single imputation methods are also continuing to increase.

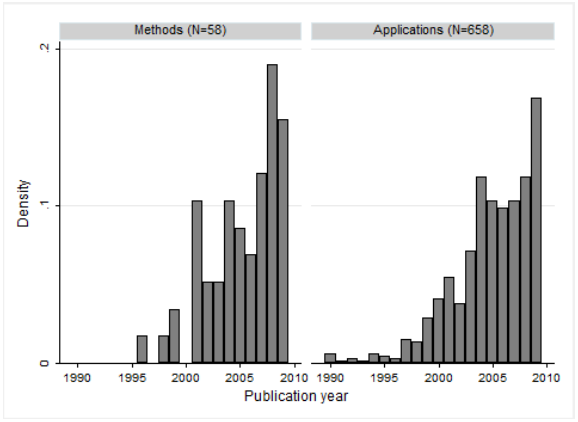


Figure 3.3: Distribution of articles reporting or describing last observation carried forward

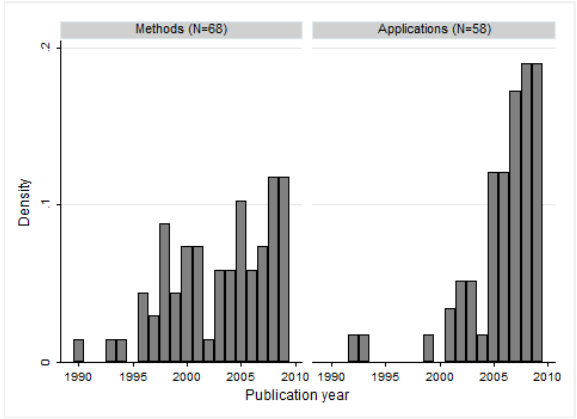


Figure 3.4: Distribution of articles reporting or describing other single imputations

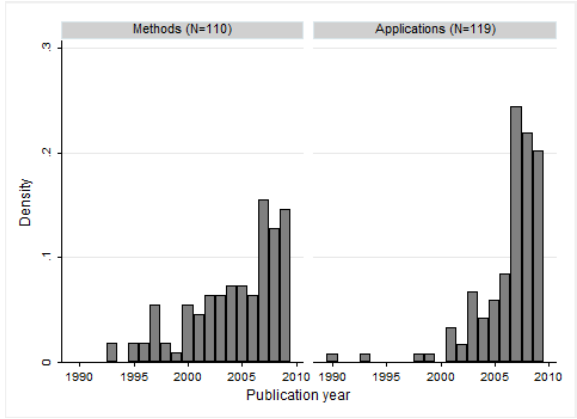


Figure 3.5: Distribution of articles reporting or describing multiple imputation

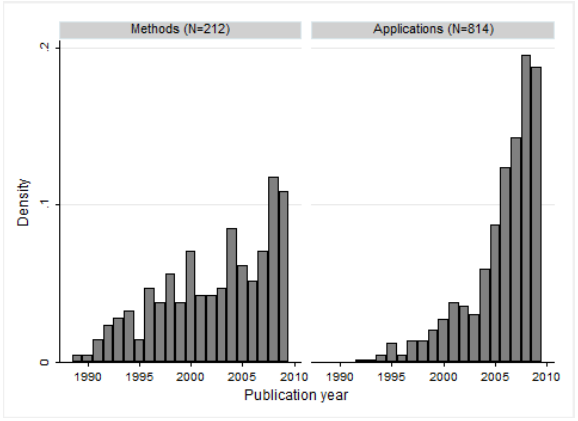


Figure 3.6: Distribution of articles reporting or describing mixed models

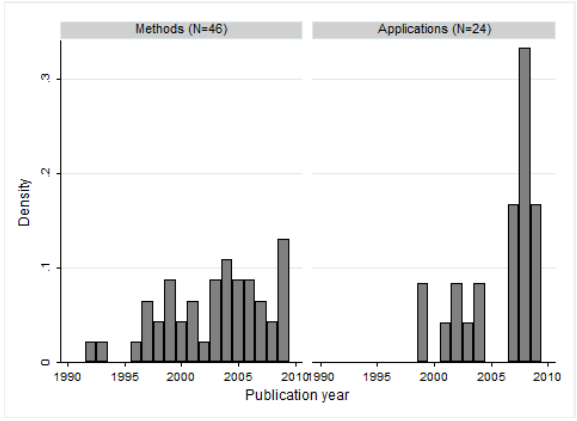


Figure 3.7: Distribution of articles reporting or describing GEEs

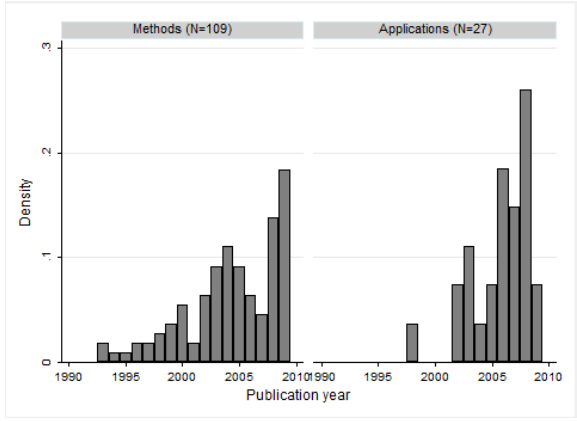


Figure 3.8: Distribution of articles reporting or describing MNAR Models

Articles which included examples of applications of missing data methods were divided into two broad groups, namely clinical trials and observational studies, according to the source of the data used in the study. Figure 3.9 displays the distribution of studies over time by source of the data and indicates similar increases in observational studies and trials.

While the number of studies applying methods increased rapidly since 2000, this could potentially be due to an overall increase in publishing rates. As it was not feasible to include all longitudinal studies in this review, a search of all articles published since 1987 and indexed in MEDLINE was conducted to provide crude estimates of overall publication rates. In total 68417 observational studies and 336268 randomised controlled trials were identified. The distributions of these studies by publication year is provided in Figure 3.10. While the proportion of studies reporting missing data methods grew exponentially, with almost 20% of trials and observational studies which used at least one method being published in 2009, the rate of increase in the total number published was less steep. Publication of observational studies increased at a steady rate, before increasing more rapidly after 2000. However, 2009 studies made up less than 15% of the total. The increase in publication of trials followed a linear trend increasing at a slower rate than the increase observed in studies applying a missing data method. This suggests that the number of studies applying one or more method as a proportion of the total number published has increased over time.

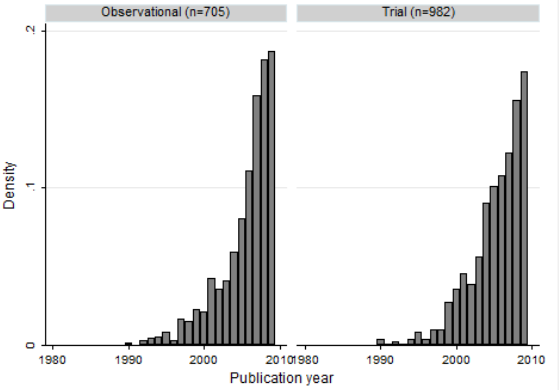


Figure 3.9: Distribution of observational studies and trials reporting applying missing data methods

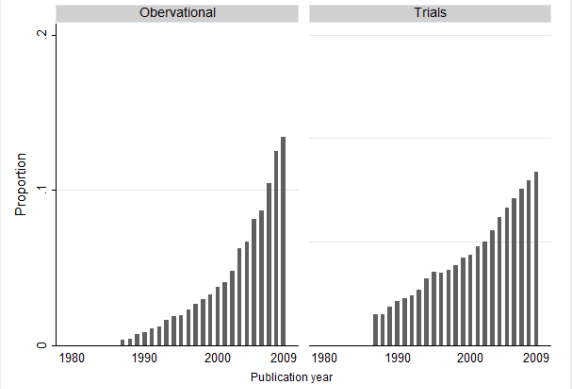


Figure 3.10: Distribution of all observational studies and trials indexed in MEDLINE

The distribution of articles by missing data method and source is presented in Figures 3.11 to 3.16. Few observational studies used last observation carried forward, but those that did apply this technique were published towards the end of the study period, whereas, its use has continued to increase in clinical trials. The earliest studies applying multiple imputation were all observational, however, its use is now increasing in both trials and observational studies.

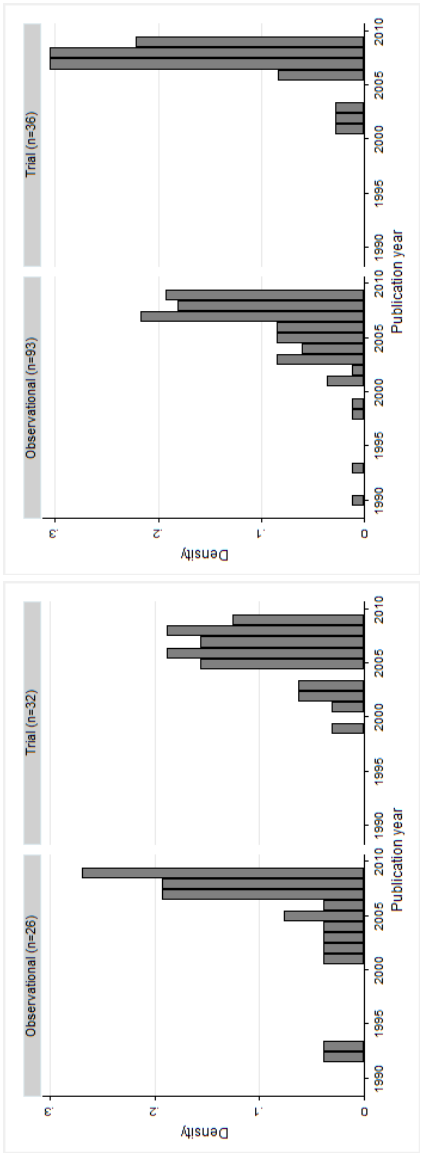


Figure 3.11: Distribution of trials and observational studies applying last observation carried forward

Figure 3.12: Distribution of trials and observational studies applying other single imputations

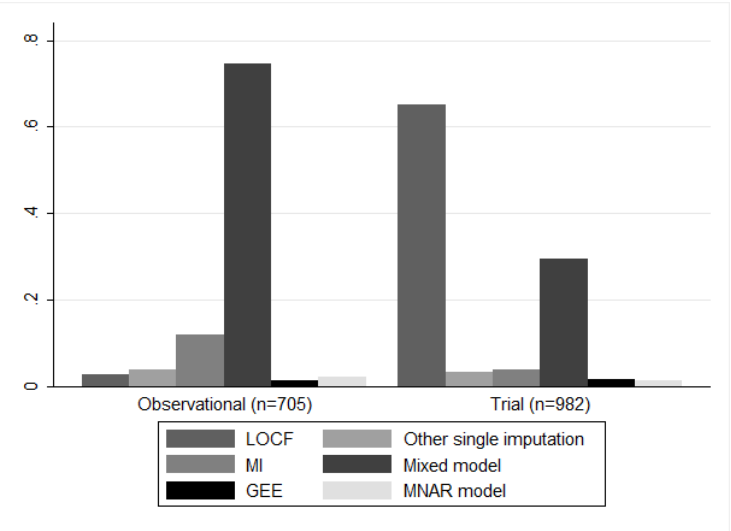
Figure 3.13: Distribution of trials and observational studies applying multiple imputation

Figure 3.14: Distribution of trials and observational studies applying mixed models

Figure 3.15: Distribution of trials and observational studies applying generalised estimating equations

Figure 3.16: Distribution of trials and observational studies applying MNAR models

Figure 3.17 illustrates the proportion of articles from trials and observational studies which applied each of the included methods. While last observation carried forward was applied in around 65% of trials, it was used by fewer than 5% of observational studies. Meanwhile, mixed-effects models were the most common method in observational studies, but they featured far less in trials.



Abbreviations: LOCF last observation carried forward, MI Multiple imputation, GEE generalised estimating equation, MNAR missing not at random

Figure 3.17: Comparison of methods used in trials and observational studies

3.7 Discussion

The articles included in this review highlight both the increased attention given to missing data in the methodological literature and the increasing use of the methods in the analysis of longitudinal studies.

The use of multiple imputation, mixed models and generalised estimating equations were all observed to rapidly increase after 2005. This is likely aided in part by the availability of routines to implement these methods in most statistical soft-

ware packages by this time. While it is encouraging to see these increases, concerns have been highlighted about the quality of reporting of multiple imputation making it impossible to determine whether the method has been correctly applied [156]. While multiple imputation can reduce bias and improve efficiency of analyses, when the imputation model is incorrectly specified they can produce poorer results than a simple complete case analysis [74, 145]. It is therefore important to ensure that methods applied to missing data are fully reported to allow readers to assess the appropriateness of the analyses for themselves.

Last observation carried forward and other single imputation methods were also observed to be increasingly applied, particularly in the analysis of clinical trials. Only 6% of the methodological literature identified discussed the use of last observation carried forward and the majority of these highlighted its limitations and the potential bias associated with its use [88, 89, 94]. Instead they suggested other methods, such as mixed models or generalised estimating equations, might be more appropriate. It was therefore surprising to observe an increase in its use, even after other methods became easily accessible.

Models suitable for data MNAR were described in 13% of the identified methodological studies. However, very few applications of these models were observed. This may be in part a result of the complexity associated with fitting these models [122] and slow integration into standard software packages. However, they can provide a useful approach to sensitivity analyses [88] when it is suspected that data may be MNAR and so further work to ensure such techniques are accessible could help integrate them into routine analyses.

It was not possible to include all longitudinal studies with missing data in the analysis to examine trends in the use of complete or available case analyses due to the volume of such studies published. Crude estimates of the number of clinical trials

and observational studies showed that the number of studies reporting an application of missing data methods has increased at a more rapid rate than the overall publication rate. This implies that the proportion of all studies applying a method is increasing over time. This is encouraging and suggests that the recently increased focus in the methodological literature on missing data methods has led to greater awareness of the impact of missing data in the routine analysis of longitudinal studies.

As overall numbers of studies published has increased, the total number published each year using complete or available case analysis is also likely to be increasing. It is therefore important to ensure that researchers have sufficient information available to them to determine which methods are appropriate for use in their own analyses.

A further limitation of this study was the reliance on electronic databases to identify studies which specified the use of a missing data method. Using PubMed, Embase and MEDLINE it was not possible to conduct searches of the full text of articles, though in Web of Science full texts were searched. It is therefore likely that some articles indexed in the database, which do not mention missing data methods in the abstract, may have been missed. With increased awareness of missing data it is possible more recent studies may be more likely to mention the method of analysis in the abstract. Previous smaller reviews of the use of missing data methods, which used detailed searches of specific journals, reported trends similar to those observed in this study [153–156], suggesting the impact of any change in the likelihood of missing data methods being reported in the abstract is likely to be small.

The work presented here was carried out in 2010-11 with articles published until the end of 2009 included. Since this time, access to methods and software for handling missing data have become more widely available. Further, the number of studies published every year was shown to be growing exponentially. Therefore ex-

panding the review to take account of new articles would not be feasible due to the expected volume of recent studies. Despite the increasing availability of methods, a more recent review of trials published in the BMJ, JAMA, Lancet and the New England Journal of Medicine in between July and December 2013 revealed that the majority of trials with missing outcome data continue to rely on complete case or single imputation methods [6]. Model based approaches were used by just 19% of 73 trials reporting missing data and multiple imputation by 8%. Sensitivity analysis was applied in 35% of the trials, but in almost all cases the assumption about the missing data mechanism underpinning the sensitivity analysis was the same as the main analysis.

A review of the reporting and handling of missing data in articles published in three leading epidemiology journals in 2010 revealed that complete case analysis was by far the most common method of analysis, used in 81% of the studies [7]. Although this review did not distinguish between cross-sectional and longitudinal studies, it revealed that as in RCTs, the use of complete case and single imputation methods were still prominent in epidemiological studies. Fourteen percent of studies used single imputation methods, 11% reported a sensitivity analysis and multiple imputation, maximum likelihood estimation and inverse probability weighting were reported by 8%, 2% and 3% of studies respectively.

Studies in which multiple methods are applied and compared across a range of scenarios can be useful in illustrating the strengths and weaknesses of different methods in different situations. Many existing comparative studies focus on continuous missing data making choice of method more difficult when missing data are binary or categorical, despite such data being common in medical research. Further, many studies that do exist use simulated data and tend to represent studies that are very simplistic in design and lack the complexity often seen in real datasets. Simulation studies mean that the performance of methods can be assessed by comparing results

to a known true value. Comparative studies that use real datasets are only able to compare results of methods to each other and can't assess bias as 'true' values are unknown. The remainder of this thesis focuses on the comparison of methods for non-continuous outcomes using a combination of real data and simulations with the aim of illustrating the impact of missing data and the performance of methods when applied to a large longitudinal cohort study.

Chapter 4

Missing data in the South London Stroke Register

4.1 Abstract

Background: The South London Stroke Register (SLSR) is a population based register of all first ever strokes in a defined area of South London. Participants are followed up three months and annually after stroke but typically only two thirds of those eligible complete each follow-up. The patterns, predictors and impact of the resulting missing data have not been fully explored.

Methods: In this chapter the data collection tools and methods used by the SLSR are described in detail. Exploratory analyses were then carried out to identify and describe patterns and predictors of missing follow-up data.

Results: Non participation at follow-up was not completely random with age and stroke severity both associated with incomplete follow-ups. Disability and activity level after stroke were lowest in those dropping out soonest and deteriorated rapidly in the follow-ups prior to dropout or death. Dropout was also more likely closer to death.

Conclusions: It is highly likely that current disability and activity levels at a given follow-up are associated with likelihood of participation and it is possible that the

missing follow-up data are *missing not at random* (MNAR).

4.2 Introduction

The South London Stroke Register (SLSR) is a longitudinal population based study collecting information on all first in a lifetime strokes in a defined area of south London. Participants are followed up at three months, one year and then annually after stroke. At any given time point, between 30 and 40% of participants who are still alive do not complete the scheduled follow-up.

Over 200 papers using SLSR data have been published with many focusing on outcomes after stroke such as cognitive impairment [157–159], incontinence [160, 161], anxiety [158], depression [158, 162], disability [158], inactivity [158], health related quality of life [163], recovery [164], epilepsy [165], prevalence of other known risk factors for stroke [166] and risk factor management [167]. Almost all of the SLSR studies on stroke outcomes have conducted an available case analysis with missing data not taken into consideration. Many do report comparisons between the baseline characteristics of those who had incomplete data and those who did not to identify where potential biases may arise. However, the extent to which the incomplete follow-up may be biasing findings has not been explored, nor has the ability of existing methods for handling missing data to correct for such biases.

To address this, in this thesis, two studies were conducted to investigate the potential impact of missing follow-up data and to compare the performance of methods for handling the missingness; the methods and results are presented in chapters 5 to 7. To inform the design of these studies exploratory analysis of the SLSR was carried out first to meet the second objective of this thesis. In the first half of this chapter an overview of the methods used and data collected by the SLSR is presented. This is followed by results of exploratory analyses of the SLSR dataset focusing on patterns and predictors of incomplete follow-up.

4.3 Background: definition and impact of stroke

Stroke is defined by the World Health Organisation (WHO) as *"rapidly developing clinical signs of focal or global neurological deficit, lasting more than 24 hours or leading to death, with no apparent cause other than of a vascular origin"* [168]. Around 80-85% of strokes are ischaemic strokes which are a result of a blockage in the blood supply to the brain [169,170]. In other cases primary intracerebral haemorrhage [PICH] causes a bleed within the brain; these cases account for around 10% of all strokes. The remaining 5% are subarchnoid haemorrhages [SAH] resulting from a bleed from a vessel just outside the brain [169,170]. More severe strokes and greater mortality rates are associated with PICH [171,172] and SAH [172] compared with ischaemic stroke.

Worldwide in 2010, around 17 million people had a first ever stroke, 6 million died from stroke and over 33 million were estimated to have survived a stroke sometime in the past [173]. Compared to 1990, these were increases in absolute numbers of 68%, 26% and 84%, respectively [173]. Globally, the overall incidence of stroke has remained stable over the past 20 years while mortality rates have decreased, with the increases in absolute numbers resulting from shifts in the demographic structure of the population [173]. Significant differences exist in trends observed in high compared to middle and low income countries. The incidence was greatest in middle and low income countries and increased, though non-significantly, by 12% (95% confidence interval -3 to 22%) between 1990 and 2010. In high income countries during the same time period incidence decreased by 12% (6-17%) [173]. However, with an ageing population and improved survival after stroke, even with this decrease in incidence the prevalence rate rose by 27% (14-38%) in high income countries [173].

In the United Kingdom [UK], stroke was the second leading cause of death in 2010, accounting for 8.8% of all deaths [174]. Here, the treatment of stroke accounts for around 5% of the total NHS health budget. When treatment and loss of pro-

ductivity are combined, it has been estimated that stroke costs the UK 8.9 billion annually [175]. Accurate estimation of the burden of stroke is important to enable efficient planning of healthcare budgets and services. Such estimations rely on data from a variety of sources with unbiased estimates from population samples providing a vital contribution when assessing the burden of stroke worldwide [158, 176].

The SLSR is one such population based study and has recorded all strokes in a defined geographical region continuously since 1995. Data collected has been used to demonstrate a decrease in stroke incidence in the UK of approximately 40% between 1995 and 2010 [177]. The mortality rate after stroke also reduced by 40% when data from participants with a stroke in 2007 to 2010 were compared to that from 1995 to 1998 [172].

The inner city position and source population of the SLSR, which consists of approximately 56% white, 25% black and 19% other ethnic groups, has enabled researchers to focus on differences between white and black populations. An early analysis of data collected between 1995 and 1996 found those from black ethnic groups were over twice as likely to have a stroke compared to white [178]. When trends in incidence over time were broken down by ethnicity it was also found that between 1995 and 2010 the decrease in incidence was significant in whites only [177]. While people of black ethnic origin were most likely to have a stroke, they have also been shown to be more likely to survive, even after adjusting for age and other confounding factors [172, 179]. Other findings from the SLSR have highlighted ethnic differences in stroke risk factors [180], type of stroke [180] and in access to acute and rehab stroke care [181].

Participants in the SLSR are followed up annually after stroke and this data has been used to contribute to the understanding of the incidence, prevalence, and predictors of adverse outcomes after stroke, including recurrent stroke [182], cognitive

impairment [157–159], incontinence [160, 161], epilepsy [165],anxiety [158], depression [158, 162], disability [158] and inactivity [158].

While the population based nature and length of both recruitment and follow-up are obvious strengths of the SLSR, one of the biggest potential limitations is the possibility of bias being introduced due to participants being lost to follow-up. Approximately 40% of follow-ups are missed by participants still alive at that point. Much of the work described above has used data from participants who have complete information; there is therefore a need to determine whether the incomplete data leads to bias and, if so, what the best method of dealing with this is. The following section describes the methods used to maximise completeness of data and summarises information collected by the SLSR, with the issues surrounding missing data described in detail in section 4.5.

4.4 SLSR data collection methods and tools

4.4.1 Source population and identification of participants

The area covered by the South London Stroke Register consists of 22 wards in the North of the boroughs of Lambeth and Southwark in South London (Figure 4.1). The population of the area is ethnically diverse. At the time of the 2011 census the total population was 357,308 individuals of which 56% were of white ethnicity, 25% black (14% black African, 7% black Caribbean, 4 % black Other), 6% Asian and 12% other ethnic groups (according to the Office of National Statistics census definitions).

Figure 4.1: SLSR catchment area

Participants are eligible to be included in the SLSR if they meet the following broad criteria:

- First in a lifetime stroke since 1st January 1995.
- Usual residence at the time of stroke is within the SLSR area.
- Stroke confirmed using WHO definition [183].

To ensure that all cases of stroke in the area are identified, multiple overlapping sources of notification are used. Using capture-recapture models, it has been estimated that between 75 and 84% of stroke cases in the area are identified by the register [184].

4.4.2 Baseline data collection

The SLSR uses specially trained fieldworkers to collect data at baseline, as soon as possible after the initial stroke event. Data collected includes demographic characteristics, prior to stroke risk factors and prescribed medications, stroke symptoms, severity and subtype, location and type of care received, newly diagnosed risk factors and medication use at discharge. The list below provides a summary of the data collected at baseline which is used in the analyses described later in this thesis. These factors were selected as previous SLSR studies have identified them as likely to be associated with completeness of follow-up and/or outcomes after stroke.

- Age at stroke onset
- Ethnicity - self-reported ethnic group using 2001 census definitions and categorised as white, black, other or unknown in the remainder of the thesis
- Gender
- Stroke Severity - measured using the Glasgow Coma Score [GCS] (where higher scores represent greater degrees of consciousness) and Barthel index (at 7-10 days after stroke)
- Stroke Subtype - defined using WHO criteria [183] and classified as ischaemic stroke, primary intracerebral haemorrhage [PICH], subarachnoid haemorrhage [SAH] or unknown/unclassified [178, 184].

4.4.3 Follow-up data collection

All surviving participants are followed up three months and annually after stroke. Trained fieldworkers conduct face to face interviews with participants where possible. The majority of data are collected using face to face interviews but when that is not possible, a postal version of the questionnaire is used instead. The most common reasons for the use of postal questionnaires are that the participant has moved out of the area or that it is more convenient, particularly for younger participants

who may be employed.

At all follow-ups information is collected relating to: service use (for example, contact with a general practitioner or specialist doctor and continuation of rehabilitation therapies), newly diagnosed risk factors and prescribed medications and other outcomes measured using standardised scales, each of which are summarised below.

4.4.3.1 The Barthel Index

The Barthel Index [BI] assesses a participants' ability to carry out activities of daily living [185] and has been included in all SLSR baseline and follow-up data collection forms since the register began. It consists of 10 items each assigned a score from 0 to 1, 2 or 3. These are then summed to give a total score ranging from 0 to 20. The items included in the BI look at whether assistance is required with walking, dressing, climbing stairs, bathing, grooming, toilet use, transfers (for example, from bed to chair) and feeding, or whether there is any faecal or urinary incontinence.

Higher scores on the BI indicate greater levels of independence in carrying out of activities of daily living, with 20 indicating full independence. The BI is often categorised for analysis grouping participants by degree of disability in the following way:

- BI = 20: Independent
- BI = 15-19: Mild disability
- BI = 10-14: Moderate disability
- BI = 0-9: Severe disability [186]

4.4.3.2 The Frenchay Activities Index

The Frenchay Activities index [FAI] is a measure of extended activities of daily living [187] and has been included in all SLSR follow-up forms. While the BI measures what a participant is capable of doing, the FAI assesses what they actually do. A score, ranging from 0 to 3, is assigned to each of 15 items according to the frequency with which an activity has been undertaken in the previous three or six months.

The activities considered cover domestic chores (such as preparing meals and washing clothes), leisure/work activities (such as reading a book) and outdoor activities (including travel outings and walking outside). Scores from each item are summed to give a total score ranging from 0 to 45 with higher scores indicating greater activity levels. The FAI can be used to categorise participants according to activity level as:

- FAI = 0-15: Inactive
- FAI = 16-30: Moderately active
- FAI = 31-45: Active [188]

4.4.3.3 Hospital Anxiety and Depression Scale

The Hospital Anxiety and Depression Scale [HADS] is a 14 item screening tool originally designed as a screening tool for use in hospital patients [189] but it has also been validated for use in stroke patients [190]. Each item on the scale forms part of either an anxiety [HADS-A] or depression [HADS-D] domain. Items are scored from 0 to 3 and are summed within each domain to give a score from 0 to 21.

A review of studies applying the HADS found a score of more than 7 in the corresponding domain to be the optimal cut-off for identifying probable cases of anxiety or of depression [191].

The HADS was added to SLSR follow-up data collection forms in 1997 and has been used in all follow-ups since that time.

4.4.3.4 Abbreviated mental test and Mini-mental state exam

Level of cognition is recorded in the SLSR at both baseline and follow-up. Up until 1st January 2000 all assessments were carried out using the mini-mental state exam [MMSE] [192] and since then using the abbreviated mental test [AMT] [193].

The MMSE is scored on a 30 point scale and takes up to 10 minutes to complete [194] while the AMT is much shorter, consisting of 10 simple questions each awarded 1 point if answered correctly [193]. Low scores on either scale can be used to identify participants with cognitive impairment; SLSR participants with an AMT <8 [195] or an MMSE <24 [196] are classed as being cognitively impaired.

4.4.3.5 SF-12 and SF-36

Health related quality of life [HRQoL] is assessed at follow-up in the SLSR. Prior to 1st March 1999, the SF-36 (36-Item Short Form Health Survey) [197] was used before switching to the shorter SF-12 [198]. The SF-12 or corresponding 12 items from the SF-36 can be used to produce a mental health and a physical health summary score [197–199]. These scores range from 0 to 100, with higher scores indicating better HRQoL. In the general population the summary scores have a mean of 50 and a standard deviation of 10.

4.4.3.6 Outcomes considered in this thesis

It is widely recognised and acknowledged that categorization of continuous data leads to a loss of information and statistical power with the loss most severe when data are dichotomised [42–44]. As described above, many of the outcomes used by the SLSR are commonly categorised for analysis [43]. Other outcomes, such as recurrent stroke events, prevalence of risk factors, speech problems, incontinence, use

of services, social support and perceived recovery are categorical in nature. Only the SF-12 is routinely analysed in its continuous form [198]. As the vast majority of studies of stroke outcome use either truly categorical variables or categorise scales prior to analysis, the focus of this thesis is on handling non-continuous missing data.

In the studies described in the following three chapters, four outcomes were considered; these were disability derived from the BI, activity level derived from the FAI and the presence of anxiety or of depression defined using the HADS-A and HADS-D, respectively. These outcomes were selected to give a mixture of binary and ordinal outcomes. As they are all derived from an underlying scale this also gave the opportunity to compare methods when imputations were made on the continuous form and the categorical form. Along with cognition, they also represent the most commonly reported consequences of stroke.

4.4.4 Recording of deaths

Details of all participants recruited to the SLSR are sent to the Health and Social Care Information Centre (HSCIC, formerly to the Office of National Statistics) for flagging. Official date and cause of death of any participants who have passed away are then notified to the register on a regular basis. Occasionally fieldworkers are notified of a death prior to receiving data from HSCIC, in which case the date (or approximate date) of death is recorded and then updated with official records once received.

4.4.5 Ethical approval

The ethics committees of Guy's and St Thomas' Hospital NHS Foundation Trust, King's College Hospital Foundation trust, National Hospital for Nervous Diseases, Queen's Square Hospital, St George's Hospital and Westminster Hospital approved the study. Before being registered patients or their relatives gave written informed consent to be included on the register and followed up.

4.4.6 My role in the SLSR

I have been employed as a research assistant on the SLSR since 2006. During this time I have been responsible for data management and cleaning of the SLSR data. I regularly attend team meetings and work closely with the field workers responsible for collecting the data. I also provide support to other internal and external researchers wishing to use the data. This includes giving guidance on the completeness of data, identifying the data required to meet the research aims and provide advice regarding, or carrying out, statistical analysis of the data. I have carried out analyses or supported other non-statisticians conducting analyses of the SLSR data and co-authored around 20 articles published in peer reviewed journals or currently under review. I have also presented findings from the SLSR at national and international stroke conferences. This has provided me with a thorough understanding of the workings of the register and the issues surrounding follow-up data collection.

4.5 Missing data in the SLSR

4.5.1 Sources of missing data

4.5.1.1 Baseline data

Overall the level of missing data among information collected at the time of stroke is low. Much of the information required is routinely collected and recorded in medical notes, and fieldworkers are trained to obtain information directly from the participant or next of kin, as appropriate. Some items do have some missing data. The items considered in this thesis as potential predictors of outcome or missing data were previously listed in section 4.4.2. Among those, age and gender have no missing data. Other key variables in the register, such as date of stroke, also have no missing data.

Ethnicity is missing for approximately 3% of participants. In the majority of these cases it was not possible to meet the participant while in hospital and the information was not recorded in medical records, and that they would not have taken part in a follow-up interview where the missing data could have been collected retrospectively.

Stroke subtype is unknown for just under 10% of the register. To be included in the register a stroke must have been confirmed; thus in these participants a stroke has occurred but it was not possible to determine the exact subtype of stroke. Therefore, unknown or unclassified subtype is considered to represent a distinct group of participants and is not treated as missing.

BI at 7-10 days is missing for approximately one quarter of the SLSR participants. However, the majority of these are participants who died shortly after stroke and so are not included in most analyses of stroke outcome. Among those with at least one follow-up, only 5% have missing data. Glasgow coma score, also used as a marker of stroke severity, is missing in less than 5% of participants.

In the exploratory analyses presented in the remainder of this chapter, participants with missing ethnicity and stroke severity measures were excluded.

4.5.1.2 Follow-up data

Missing data patterns at follow-up are complex but typical of studies following elderly populations [2]. There are a number of reasons why any given follow-up may not have been completed. Although every effort is made to identify and recruit patients to the register as soon as possible after stroke onset, a number of patients are only notified to the register some months (and in a few extreme cases years) after onset. This may be due to the patient having had the stroke outside the study area, for example while on holiday, or they may not have been admitted to hospital. The earliest follow-up points may therefore have already been passed by the time

the patient is recruited, resulting in missed follow-ups.

At each follow-up fieldworkers are unable to contact a number of patients. This can results from patients moving home or change of telephone details. Where patients details are found to no longer be correct, or when repeated attempts to make contact fail, fieldworkers attempt to update records and trace the patients through contact with GP's, NHS tracing and by checking recent hospital records. If current contact details cannot be obtained then the process is repeated the following year. It is not uncommon for patients to miss several years of follow-ups before contact is made and they then contribute further data to the study.

A small number of patients also refuse to take part in a given follow-up but still wish to remain in the study while a small number also request that no further contact is made and they drop out of the study and are not contacted again.

The majority of missing data are a result of follow-ups not being conducted at all, but among those that are, item non-response can also be a problem. Most of the data collected at follow-up can be obtained from a relative or carer if the participant is unable to answer for themselves. The measures used to assess cognition and the HADS cannot be obtained from a proxy. Therefore if a participant is unable to, or has trouble answering directly, these scales in particular will be incomplete.

The second objective of this thesis was to describe the patterns and predictors of incomplete follow-up in the SLSR. The results of the exploratory analyses undertaken to achieve this objective are presented in the remainder of this chapter. The results were then used to inform the design of the simulation study, described in Chapter 5.

4.5.2 Completeness of follow-up data

Between 1st January 1995 and 31st December 2009 a total of 3617 first ever strokes were registered by the SLSR. Figure 4.2 describes the flow of participants through follow-up, providing the numbers included at each follow-up time point. Participants were considered 'not eligible' for a follow-up (and any future follow-ups) if the specified point had not yet been reached. Participants registered later than the date of a given follow-up are described as late notifications in the diagram. Late notification was the reason for incomplete three month follow-ups in 409 out of 2629 (15.6%) patients alive three months after stroke. Twenty-two patients were recruited at least five years after stroke, with one identified more than 11 years beyond the initial event.

The number of patients surviving to each follow-up point reduced year by year. To ensure a sufficient number of participants to explore the relationship between completeness of follow-up and baseline sociodemographic and clinical characteristics and to ensure a sufficient sample size for the studies presented later in this thesis, only the first five years of follow-up were considered in analyses from this point onwards. Further, to ensure all participants included were eligible for all five years of follow-up, analyses were restricted to participants with a date of stroke before 31st December 2005.

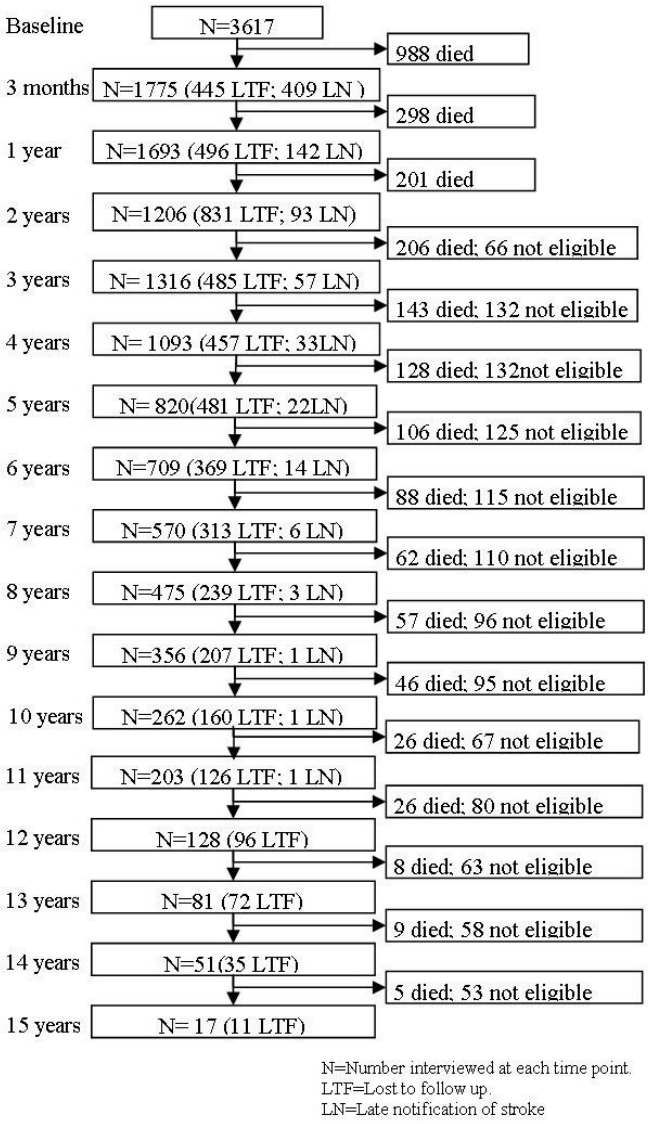


Figure 4.2: Flow of SLSR participants over 15 years of follow-up

In total 3145 participants had a stroke before the end of 2005. The status of these patients at each follow-up point is summarised in Table 4.1. Participants were defined as having dropped out if they missed a follow-up and completed no further follow-ups until death or the end of the study period. Otherwise, at least one future follow-up must have taken place and so was classed as an intermittent missed follow-up. By five years after stroke 56% of the participants had died. Among those who survived approximately 70% completed the early follow-ups, with the rate falling to just below 60% by the five year follow-up.

Table 4.1: Follow-up status of SLSR patients (1995-2005)

Follow-up	3 month	1 year	2 years	3 years	4 years	5 years
All participants, n(%)						
Follow-up complete	1595(50.7)	1472(46.8)	1293(41.1)	1155(36.7)	1040(33.1)	817(26.0)
Dropout	228(7.3)	183(5.8)	222(7.1)	217(6.9)	245(7.8)	318(10.1)
Intermittent Missed	437(13.9)	340(10.8)	318(10.1)	273(8.7)	228(8.7)	253(8.1)
Died	885(28.1)	1150(36.6)	1312(41.7)	1500(47.7)	1632(47.7)	1757(55.9)
Survivors only, n(%)						
Follow-up complete	1595(70.3)	1472(73.8)	1293(70.5)	1155(70.2)	1040(68.7)	817(58.9)
Follow-up incomplete	665(29.7)	523(26.8)	540(29.5)	490(29.8)	473(31.3)	571(41.1)

Among participants who were alive five years after stroke, so eligible for a total of six follow-ups, only 24.6% completed all six. A further 28.5% missed one follow-up and 15.2% missed two. Six percent of participants known to be alive at five years had had no follow-ups completed, and a further 6.0% had only one.

While 60-70% of survivors took part in follow-up interviews, there was also non-response to individual items or full scales which meant that even when a follow-up was done, some participants still had missing data. The rate of missing data in the BI, FAI and the anxiety and depression domains of the HADS are summarised in Table 4.2. Overall the BI and FAI are completed at the majority of follow-ups.

Where they are incomplete, there is often only a single item missing in questionnaires completed and returned by post. However, up to one quarter of participants at three months did not complete the HADS. While the missing BI and FAI data are likely to be missing completely at random, the HADS is most often due to some underlying reason which prevents the participant from being able to complete the scale. Unfortunately, the specific reason has not been routinely recorded in the SLSR, but it's likely that those who do not complete the HADS are sicker than those who do.

Table 4.2: Item non-response in the SLSR (1995-2005)

	3 month	1 year	2 years	3 years	4 years	5 years
Follow-up complete, (N)	1595	1472	1293	1155	1040	817
BI, (% missing)	0.8	0.8	0.9	3.5	4.2	5.4
FAI, (% missing)	4.6	4.1	3.3	4.2	5.1	6.2
HADS-A, (% missing)	23.2	21.9	23.2	20.2	14.8	14.4
HADs-D,(% missing)	24.3	21.3	23.5	19.9	15.0	14.0

Percentages calculated using the proportion of participants who completed a follow-up but which did provide complete data on specific scales.

Abbreviations: BI Barthel index, FAI Frenchay activities index, HADS-A hospital anxiety and depression scale - anxiety domain, HADS-D hospital anxiety and depression scale - depression domain.

4.5.3 Predictors of incomplete follow-up

Table 4.3 contains the results of analyses conducted to examine the relationship between characteristics of participants at baseline and incomplete follow-ups up to five years after stroke. Multivariable logistic regression models were used to compare the odds of completing a follow-up in participants surviving at least to the time of the follow-up in question. No associations between gender and incomplete follow-up were found. There were significant associations found with age, with the odds of

completing follow-up increasing by between 10 and 20% for every 10 year increase in age at the time of stroke. Ethnicity and stroke severity, when assessed using the GCS were not significant. BI at 7-10 days after stroke was a strong predictor of completeness of follow-up at three months with participants who were independent half as likely to complete the three month follow-up as those with moderate to severe disability. Beyond three months, BI was not significantly associated with incomplete follow-up. Type of stroke was not associated with incomplete follow-up at the earlier time points but was significant at four and five years after stroke.

Table 4.3: Associations between baseline characteristics and odds of complete follow-up among survivors

Follow-up	3 month		1 year		2 years		3 years		4 years		5 years	
	OR (95%CI)	p	OR (95%CI)	p	OR (95%CI)	p	OR (95%CI)	p	OR (95%CI)	p	OR (95%CI)	p
Female gender	0.8(0.7-1.1)	0.204	1.0(0.8-1.2)	0.865	1.1(0.9-1.3)	0.575	0.9(0.7-1.1)	0.274	1.3(1.0-1.5)	0.073	1.1(0.9-1.4)	0.268
Age (per 10 years)	1.1(1.0-1.2)	0.053	1.1(1.1-1.2)	0.001	1.2(1.1-1.3)	<0.001	1.2(1.1-1.3)	<0.001	1.2(1.1-1.3)	0.001	1.1(1.0-1.2)	0.058
Ethnicity												
White	1	0.786	1	0.423	1	0.856	1	0.259	1	0.345	1	0.150
Black	0.9(0.7-1.2)		0.9(0.7-1.2)		1.0(0.7-1.3)		0.9(0.6-1.1)		0.9(0.7-1.2)		0.8(0.6-1.0)	
Other	0.9(0.6-1.4)		0.7(0.5-1.1)		0.9(0.6-1.4)		0.7(0.5-1.1)		0.8(0.5-1.2)		0.9(0.6-1.4)	
Glasgow Coma Score	1.0(1.0-1.0)	0.466	1.0(1.0-1.0)	0.965	1.0(0.9-1.0)	0.281	1.0(0.9-1.0)	0.287	1.0(1.0-1.1)	0.854	1.0(0.9-1.1)	0.607
Barthel at 7-10 days												
M-S disability	1	<0.001	1	0.152	1	0.206	1	0.066	1	0.315	1	0.730
Mild disability	0.9(0.7-1.2)		0.8(0.6-1.1)		1.0(0.7-1.3)		0.9(0.7-1.3)		1.3(0.9-1.7)		1.1(0.8-1.5)	
Independent	0.5(0.4-0.6)		0.8(0.6-1.0)		0.8(0.6-1.0)		0.8(0.6-1.0)		1.0(0.8-1.4)		1.0(0.7-1.2)	
Stroke Subtype												
Ischaemic	1	0.109	1	0.054	1	0.760	1	0.667	1	0.001	1	0.006
PICH	0.8(0.6-1.1)		0.7(0.5-1.0)		0.9(0.7-1.3)		0.9(0.6-1.3)		1.1(0.8-1.5)		1.0(0.7-1.4)	
SAH	0.6(0.4-1.0)		0.6(0.4-1.0)		0.9(0.7-1.3)		0.7(0.5-1.2)		1.0(0.6-1.7)		0.9(0.5-1.4)	
Unknown	0.7(0.5-1.2)		0.7(0.4-1.1)		0.8(0.5-1.3)		0.9(0.6-1.4)		0.4(0.3-0.7)		0.4(0.2-0.7)	

Odds ratios were estimated from multivariable logistic regression models applied separately to each time point.

Abbreviations: OR odds ratio, CI confidence interval, p p-value, M-S moderate-severe, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

Time of death was also found to be associated with completeness of follow-up. In Table 4.4 the percentage of surviving patients followed up at each time point is presented broken down by time of death. The figures in bold represent participants who were alive at that follow-up but who had died before the next follow-up time point. At all follow-up points, the follow-up rate among those dying prior to the next follow-up was more than 10% lower than that for patients who died later in the study.

Table 4.4: SLSR follow-up rates according to time of death

Follow-up, %	N	3 month	1 year	2 years	3 years	4 years	5 years
Alive,	957	62.7	69.7	68.9	65.0	57.2	44.2
Time of death							
< 3 months	885						
3 months - 1 year	265	**59.3**					
1 - 2 years	162	70.8	**68.8**				
2 - 3 years	188	72.1	70.2	**59.1**			
3 - 4 years	132	74.5	80.0	73.1	**65.5**		
4 - 5 years	125	76.0	79.2	75.2	78.4	**64.0**	
> 5 years	431	77.7	79.4	78.7	75.9	71.9	64.7

Overall, younger participants, those with less severe stroke or with unknown stroke subtype, were the most likely to miss a follow-up, particularly at the earliest follow-up points. This could potentially be due to younger, less severe cases not being admitted to hospital or having a very short stay in hospital, and so not being seen face to face by fieldworkers at the time of stroke. They may also move home or return to work and become more difficult to contact. This may also have lead to difficulties in obtaining up to date contact details for these participants resulting in lower follow-up rates soon after stroke. However, all participants are actively pursued at each follow-up and effort is made each year to obtain updated contact information; this may partly explain why these associations diminish over time. Participants who

died soon after a follow-up point were also more likely to be missed, so it is possible that those who are sickest at the time of follow-up are actually the ones most likely to be missing.

4.5.4 Predictors of outcome after stroke

Linear regression models were applied to examine associations between outcome after stroke and the participant characteristics considered as potential predictors of completeness of follow-up in the previous section. The BI is skewed to a greater degree than any of the other outcomes and age and GCS are slightly negatively skewed. However, residuals from all models were approximately normal and so no transformations were made before fitting the models. In Table 4.5 the associations with BI, FAI, HADS-A and HADS-D at one year after stroke are presented, with the corresponding data at five years after stroke summarised in Table 4.6. Age, ethnicity and stroke severity were found to be strongly associated with outcomes at both one and five years after stroke. These were characteristics that are also associated with completeness of follow-up.

Table 4.5: Associations between baseline characteristics and outcome at one year after stroke

	Barthel Index		Frenchay		HADS - Anxiety		HADS - Depression	
	β(se)	p	β(se)	p	β(se)	p	β(se)	p
Female gender	-0.7(0.3)	0.005	0.6(0.6)	0.334	1.6(0.3)	<0.001	0.5(0.3)	0.071
Age (per 10 years)	-0.9(0.1)	<0.001	-2.4(0.2)	<0.001	-0.5(0.1)	<0.001	-0.1(0.1)	0.270
Ethnicity								
White	Ref	0.030	Ref	<0.001	Ref	<0.001	ref	0025
Black	0.4(0.3)		-0.7(0.7)		-1.5(0.4)		-0.7(0.3)	
Other	-1.2(0.5)		-4.6(1.2)		0.6(0.6)		1.1(0.6)	
Glasgow Coma Score	0.2(0.1)	<0.001	0.2(0.1)	0.030	-0.1(0.1)	0.298	-0.1(0.1)	0.08
Barthel at 7-10 days								
Mod-Severe disability	Ref	<0.001	Ref	<0.001	Ref	0.013	Ref	<0.001
Mild disability	4.1(0.3)		6.5(0.7)		-0.6(0.4)		-1.0(0.4)	
Independent	4.7(0.3)		13.1(0.7)		-1.0(0.3)		-2.1(0.3)	
Stroke Subtype								
Ischaemic	Ref	0.029	Ref	0.017	Ref	0.605	Ref	0.418
PICH	1.0(0.40		1.7(0.9)		-0.4(0.5)		-0.4(0.5)	
SAH	1.3(0.7)		4.3(1.7)		0.1(0.8)		-1.1(0.7)	
Unknown	-0.4(0.6)		1.0(1.3)		0.7(0.7)		0.2(0.7)	

β coefficients were estimated using multivariable linear regression models applied separately to each outcome.

Abbreviations: se standard error, p p-value, HADS Hospital anxiety and depression scale, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage

Table 4.6: Associations between baseline characteristics and outcome at five years after stroke

	Barthel Index		Frenchay		HADS - Anxiety		HADS - Depression	
	B(se)	p	B(se)	p	B(se)	p	B(se)	p
Female gender	-1.0(0.3)	0.004	0.4(0.9)	0.677	0.6(0.4)	0.121	-0.6(0.4)	0.079
Age (per 10 years)	-0.7(0.1)	<0.001	-3.0(0.4)	<0.001	-0.6(0.2)	<0.001	-0.2(0.1)	0225
Ethnicity								
White	Ref	0.255	Ref	0.005	Ref	0.068	Ref	0.485
Black	0.3(0.4)		-2.0(1.1)		-0.1(0.5)		-0.5(0.4)	
Other	-1.0(0.7)		-5.6(1.7)		-1.1(0.7)		-0.5(0.7)	
Unknown	1.2(1.5)		1.9(3.9)		1.8(1.6)		1.1(1.5)	
Glasgow Coma Score	0.2(0.1)	0.004	0.5(0.2)	0.018	0.1(0.1)	0.218	0.0(0.1)	0.903
Barthel at 7-10 days								
Mod-Severe disability	Ref	<0.001	Ref	<0.001	Ref	0.057	Ref	0.003
Mild disability	2.1(0.5)		3.6(1.2)		0.7(0.5)		-0.0(0.5)	
Independent	3.3(0.4)		8.1(1.1)		-0.5(0.4)		-1.3(0.4)	
Stroke Subtype								
Ischaemic	Ref	0.085	Ref	0.002	Ref	0.030	Ref	0.023
PICH	0.5(0.5)		1.6(1.4)		-1.0(0.6)		-1.0(0.6)	
SAH	1.8(0.7)		7.6(2.0)		-1.8(0.8)		-2.0(0.7)	
Unknown	0.0(0.1)		-0.3(2.6)		-1.9(1.3)		0.0(1.2)	

β coefficients were estimated using multivariable linear regression models applied separately to each outcome.

Abbreviations: se standard error, p p-value, HADS Hospital anxiety and depression scale, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage

4.5.5 Relationship between incomplete follow-up and outcome

To explore the relationship between outcome and completeness of follow-up participants were divided according to time of death, dropout or first missed follow-up (if they did not drop out). The mean BI, FAI and HADS at each follow-up were then calculated according to time of death, dropout or missed follow-up.

The mean profiles for the BI are displayed in Figures 4.3 to 4.5. In Figure 4.3, only patients with complete data prior to death are included. Scores were highest in those completing all follow-ups and, at each time point, scores tended to be lowest, or poorest, in those dying sooner. Figure 4.4 represents patients who were alive until at least five years after stroke but who had complete data until a given follow-up, after which no further data have been collected. Here, there is less variation in mean scores compared to in patients who died, however, lower scores still appear associated with earlier dropout from the study. In Figure 4.5 any patient who did not drop out but was alive five years post stroke was included. They are grouped according to the time of the first missed follow-up (although all completed at least one further follow-up) and here there appears to be less distinction between the profiles.

Similar plots exploring the association between activity level, measured using the FAI are presented in Figures 4.6 to 4.8. These figures exhibit similar patterns to those observed when considering the BI.

Mean HADS-A and HADS-D scores are displayed in Figures 4.9 to 4.11 and Figures 4.12 to 4.14, respectively. While there does not appear to be any pattern in the HADS-A profiles or HADS-D scores of those who had incomplete follow-up, mean HADS-D scores were observed to increase sharply in the follow-up prior to death (Figure 4.12).

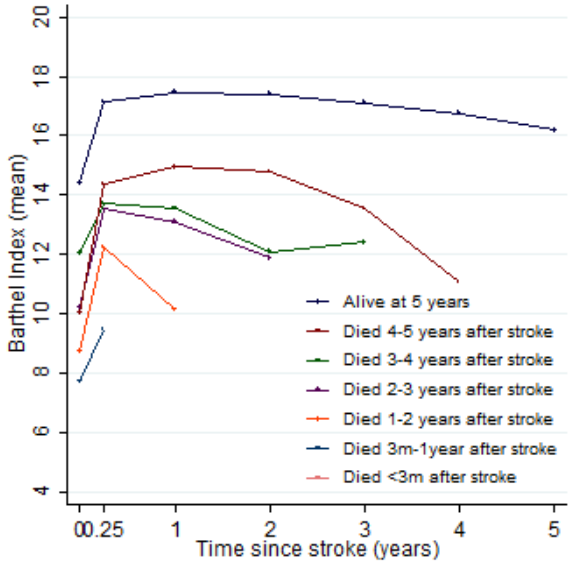


Figure 4.3: Mean Barthel Index prior to death

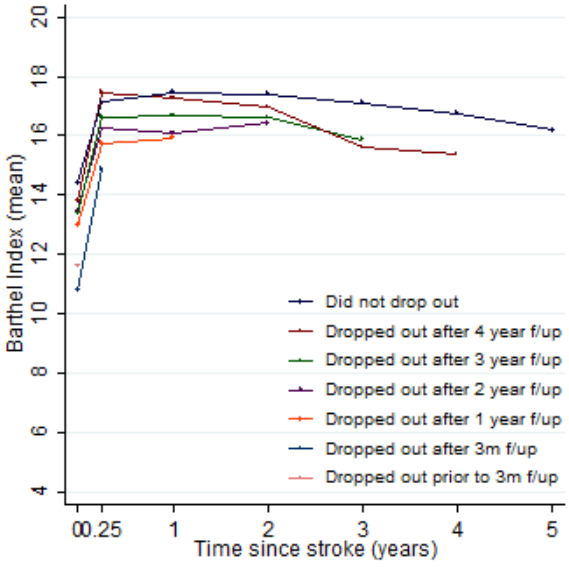


Figure 4.4: Mean Barthel Index prior to dropout in five year survivors

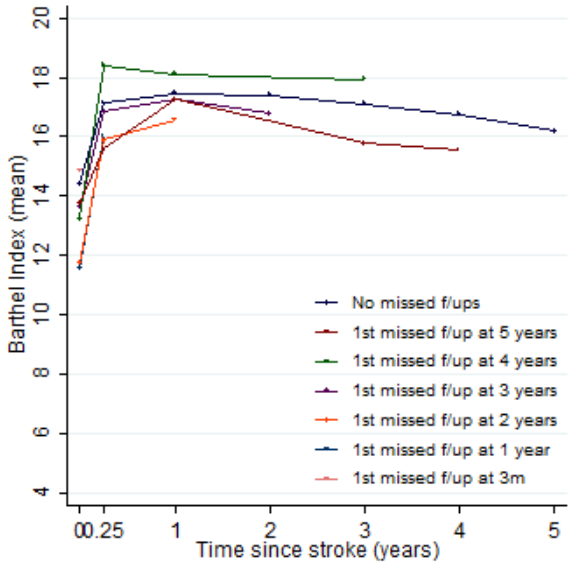


Figure 4.5: Mean Barthel Index prior to first missed follow-up in five year survivors

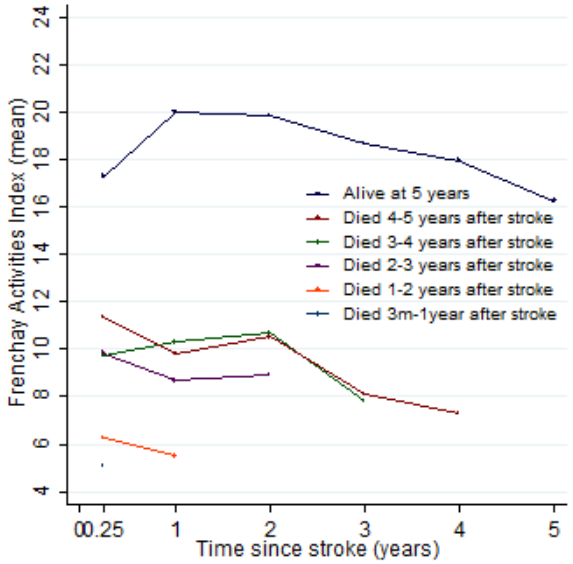


Figure 4.6: Mean Frenchay Activities Index prior to death

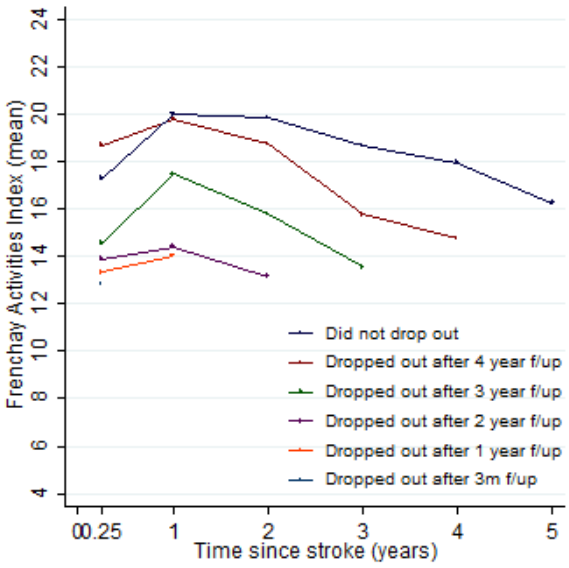


Figure 4.7: Mean Frenchay Activities Index prior to dropout in five year survivors

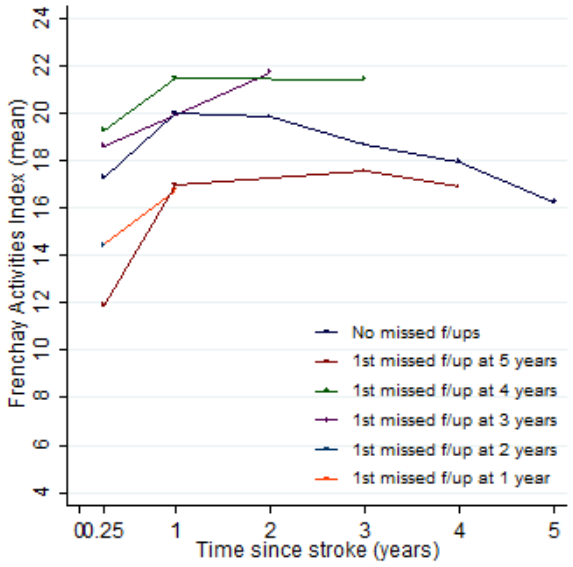


Figure 4.8: Mean Frenchay Activities Index prior to first missed follow-up in five year survivors

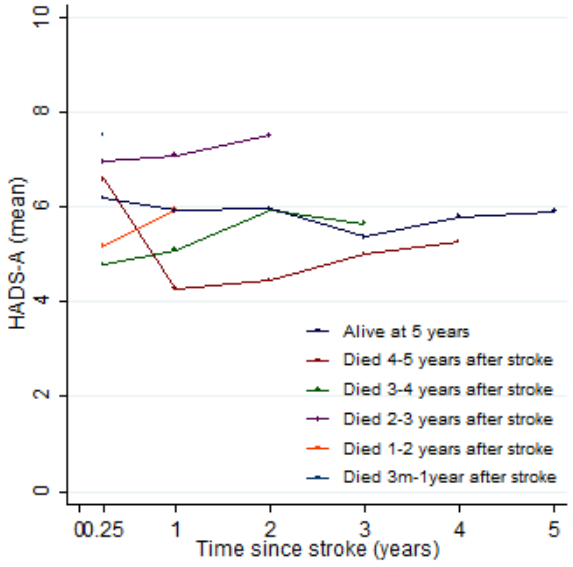


Figure 4.9: Mean anxiety score prior to death

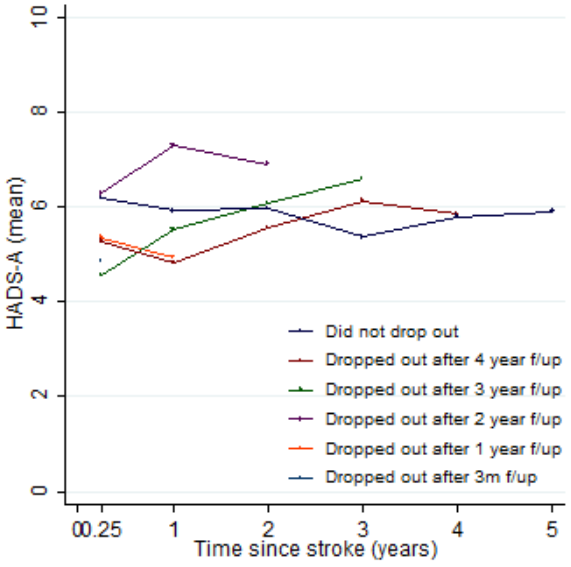


Figure 4.10: Mean anxiety score prior to dropout in five year survivors

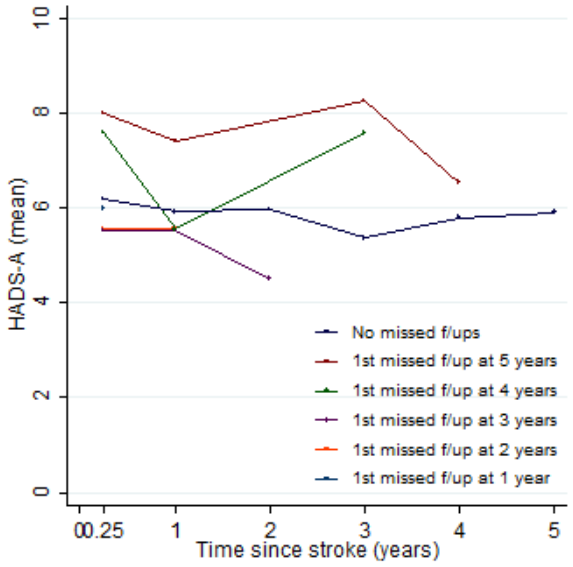


Figure 4.11: Mean anxiety score prior to first missed follow-up in five year survivors

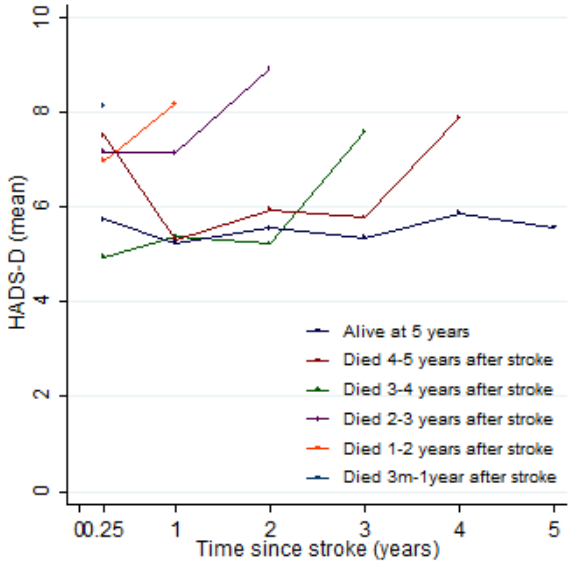


Figure 4.12: Mean depression score prior to death

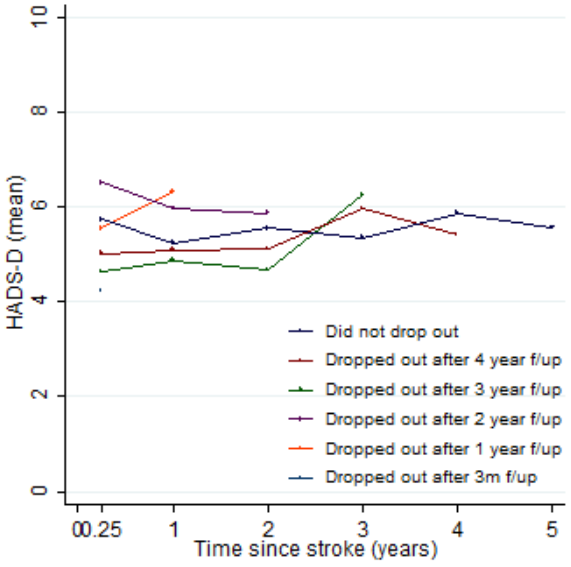


Figure 4.13: Mean depression score prior to dropout in five year survivors

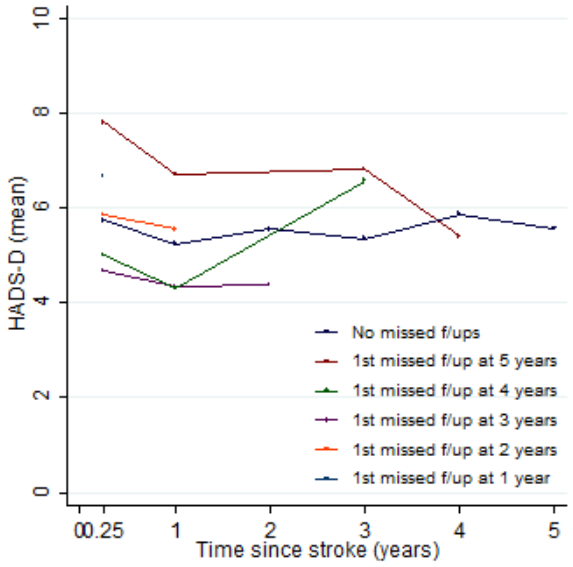


Figure 4.14: Mean depression score prior to first missed follow-up in five year survivors

4.5.6 Item non-response at follow-up

The majority of missing data in the SLSR results from non-participation in follow-up interviews. Among those who do participate item response is generally low across most items, and relatively uncommon in the BI and FAI (Table 4.2). Up to one quarter of those who participate in interviews do not complete the HADS. In most cases this is due to the participant being unable to complete the scale themselves and so they are generally sicker and older than other participants. The relationship between the outcomes after stroke and incomplete HADS measurements were explored and summarised in Figures 4.15 to 4.18. The figures represent data from all participants who completed a follow-up at five years after stroke who were then stratified into two groups representing those who did and did not complete the HADS.

The mean BI scores across all follow-ups were lower in those with incomplete HADS measurements than in those with complete data (Figure 4.15), as were FAI scores (Figure 4.16). Levels of depression were also slightly higher in those missing the HADS (Figure 4.18) but anxiety levels did not differ (Figure 4.17). Overall the trajectories of those who did not respond were similar to those who dropped out at five years.

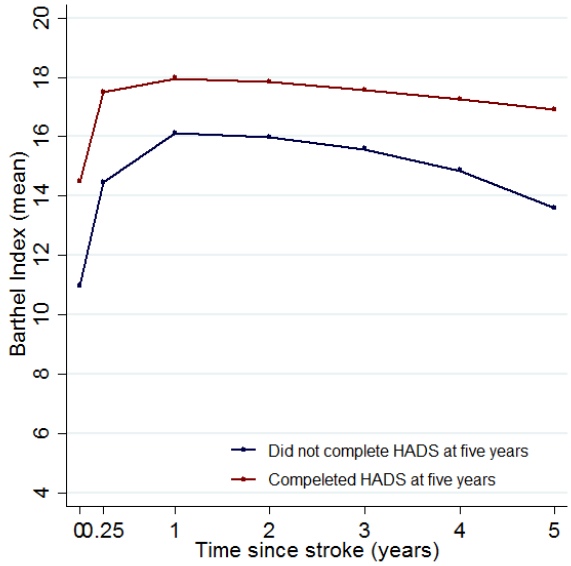


Figure 4.15: Mean Barthel Index in participants who complete the five year follow-up but did and did not complete the HADS

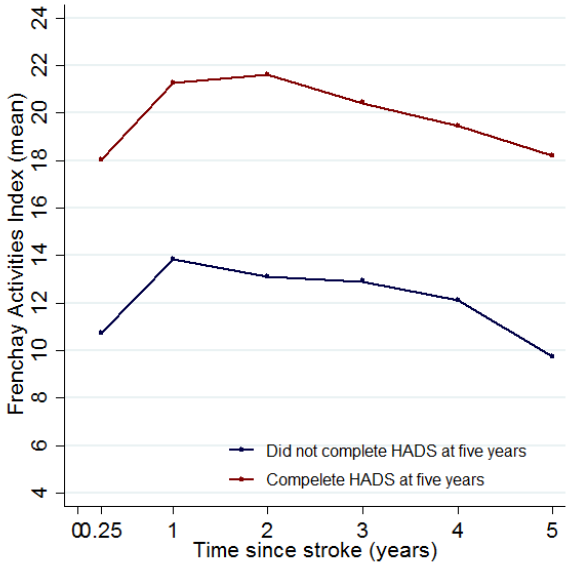


Figure 4.16: Mean Frenchay Activities Index in participants who complete the five year follow-up but did and did not complete the HADS

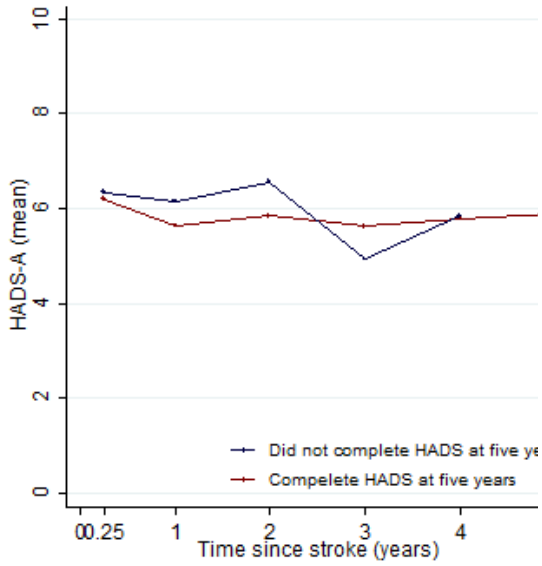


Figure 4.17: Mean anxiety score in participants who complete the five year follow-up but did and did not complete the HADS

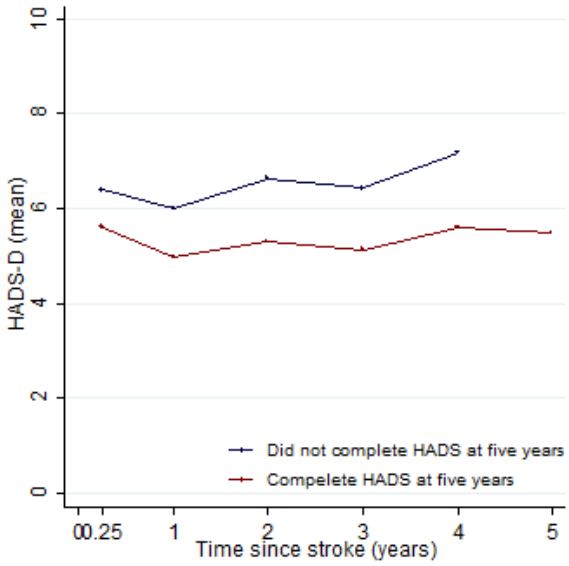


Figure 4.18: Mean depression score in participants who complete the five year follow-up but did and did not complete the HADS

4.6 Summary

The SLSR is a rich source of data and has provided valuable contributions to the study of stroke incidence, management and outcomes. While every effort is made to minimise missing data due to missed follow-up interviews, often this is unavoidable. Exploratory analyses of the SLSR dataset identified younger participants with less severe strokes to be the most likely to miss a follow-up when associations between baseline data and individual follow-ups were examined. Meanwhile, participants with similar characteristics at the time of stroke were also those who had the best outcomes after stroke. Where the scales considered in these analyses are dichotomised to enable estimation of the prevalence of poor outcome in any of the domains, there is the potential for bias if only those with complete data are considered.

The issue of missing data in the SLSR is further complicated by the relationship between completeness of follow-up and outcome. BI scores were found to be lowest among participants who dropped out soonest after stroke. There was also shown to be a decline in health (in particular in the BI and HADS-D scores showing a sharp decline) prior to death while follow-ups were more likely to be missed in the years immediately prior to death.

Based on the observed associations between completeness of follow up and it is plausible that at least some of the data are MNAR. An increase in the rate of decline of BI appears to be associated with drop-out. For BI to be considered MAR data prior to drop out would need to be fully observed and an assumption made that drop out depends only on the decline in disability level up to one year prior to the drop out occurring. In reality, before dropping out, many participants miss one or more follow up and so changes between previous follow ups are not fully observed. It is also plausible that there would be even steeper declines in in the year during which drop out occurred and that current BI may still be associated with drop out even after adjusting for previous changes. Similarly FAI showed similar patterns to,

and is highly correlated with, BI and so may also be MNAR.

Anxiety and depression had less of an association with drop out or intermittently missing data and could potentially be considered MAR. Unlike disability level which tends to decline steadily after an initial recovery period [200,201], symptoms of both anxiety and depression tend to vary over time [162, 202] and so it may be follow-ups are missed during an episode of high depressive symptoms or increased anxiety. Therefore a MNAR process dependent on current anxiety or depression levels cannot be ruled out. Therefore a MNAR process dependent on current anxiety or depression levels cannot be ruled out.

To explore further the impact that the incomplete follow-up information may have had on estimating prevalence of poor outcome and on exploring association with baseline characteristics, two separate studies were conducted. The first, a simulation study, was designed to reflect the patterns of missing data described in this chapter as closely as possible. The methods and results of these studies are presented in the following three chapters.

Chapter 5

Analysis methods

5.1 Introduction

The patterns of missing data observed in the South London Stroke Register (SLSR) were described and illustrated in the previous chapter. While it is clear that certain groups of stroke survivors are at greater risk of missing follow-ups, it is not known what consequences this may have on analyses focusing on outcome after stroke.

Two studies were conducted to examine the impact of missing data on the performance of missing data methods when estimating the prevalence and predictors of poor outcome after stroke. The results of these studies, presented in Chapters 6 and 7, will be used to address the third objective of this thesis which was to compare and determine the most appropriate methods for handling missing data in the SLSR. The objective was further broken into three sub-objectives. These were:

a To compare the performance of missing data methods when estimating prevalence of poor outcome.

b To compare results of analyses of non-continuous outcomes derived from a continuous measure when imputation techniques are applied before and after transformation.

c To compare the performance of missing data methods when identifying predictors of poor outcome.

The first study addressed objectives a and b, while objective c was explored in the second and the methods used in both studies are described in the remainder of this chapter.

5.2 Study 1: Simulation study comparing missing data methods for estimating prevalence of poor outcome after stroke

5.2.1 Overview of method

A brief overview of the process used to carry out the simulation study is described below and summarised in Figure 5.1. A subset of the SLSR in which all participants had complete follow-up data was used. From this dataset, data from a random sample were extracted in order to obtain a complete data sample in which the ratio of survivors to deaths at each follow-up point reflected that in the SLSR. This dataset was then analysed to obtain the prevalence of poor outcome when there is no missing data present. These estimates were saved to later be compared to estimates obtained after missing data were introduced. For the purpose of this thesis, poor outcome was defined as moderate to severe disability (Barthel Index <15), inactivity (Frenchay Activities Index <16) and anxiety and depression (defined as a score >7 on the relevant sub-scales of the Hospital Anxiety and Depression Scale).

Next, missing data were then generated in the dataset by removing a random sample of outcome data according to one of four scenarios. The four different scenarios used varying probabilities of dropout and intermittently missed follow-ups to produce missing data that was missing completely at random (MCAR), missing at random (MAR) dependent on baseline characteristics only, MAR dependent on baseline

characteristics and time of death and missing not at random (MNAR) depending depending on current disability level. The fact data within the sample are complete, make it possible to simulate missing data where by the missingness depends on health status at the point at which the data are missing. Further, as the data used are longitudinal and the previous chapter has shown that people tend become more unwell over time, by creating a drop out process in which those who are most disabled at a given time point are most likely to drop out, then at both that point and future time points it will be those who are most unwell who are most likely to be missing.

Within the dataset with now incompplete outcome measures, a number of different methods for analysing data in the presence of missing data were then applied and parameter estimates saved. The process was then repeated 1000 times for each of the four missing data scenarios.

Parameter estimates from missing data analyses were then compared to the 'true' values obtained earlier from analyses carried out on the datasets without missing values. Two performance measures (bias and precision, described in section 5.2.5) were then derived by averaging across all methods of analysis and missing data scenarios. In the following sections each of these stages is described in more detail.

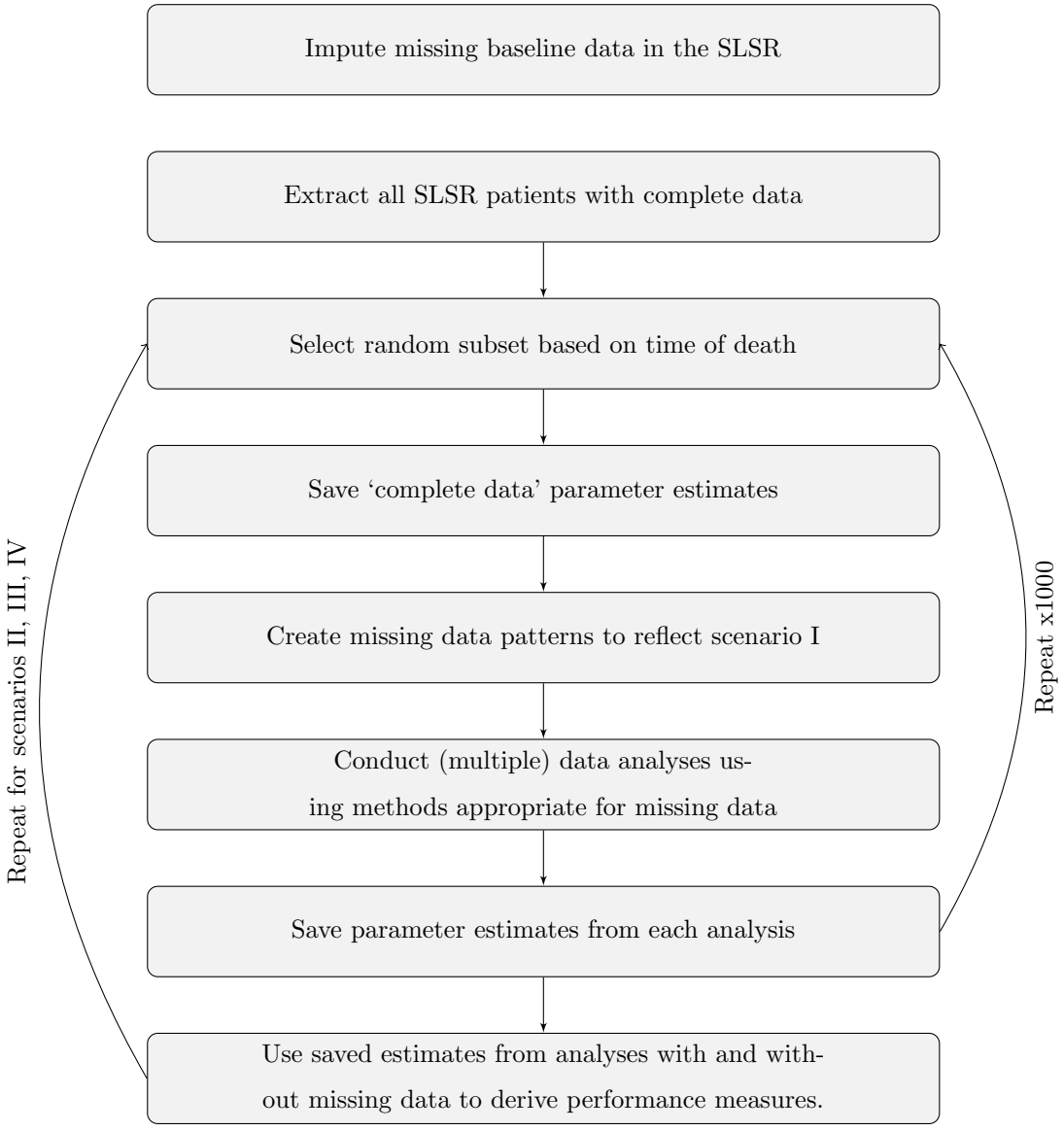


Figure 5.1: Flow chart illustrating process used to conduct simulation studies

.

5.2.2 Software

All simulations were carried out in Rv2.12 [203]. Details of any packages used are given where applicable in the following sections. The full R code used to run the simulations is available with details provided in Appendix A.

5.2.3 Generation of simulation datasets

To be eligible for inclusion in the SLSR complete data subset, participants were required to have had a stroke between 1995 and 2005, and so eligible for at least five years of follow-up at the time of data extraction in 2011. It was also required that they had outcome data recorded across every follow-up, up to five years after stroke, or where they had died before this point, complete data at every follow-up prior to death. The outcome measures required to be complete for inclusion were those used to define poor outcome (i.e. Barthel index [BI], Frenchay activity index [FAI] and the Hospital Anxiety and Depression scale [HADS]).

5.2.3.1 Missing baseline variables

In order to simulate data that is MAR and MNAR, the probability of each participant in the complete dataset dropping out or missing a follow-up was derived from the whole SLSR. This was achieved by fitting models to the SLSR to predict missingness based on baseline characteristics and time of death (see Sections 5.2.3.5 and 5.2.3.6 for details). It was therefore desirable that the baseline data selected for inclusion in the prediction models was complete for all SLSR participants.

Baseline variables, collected across all patients since 1995, which are known or expected to be associated with missingness or poor outcome after stroke and so were utilised in the simulation studies were:

- Age at stroke onset
- Gender

- Ethnicity (white, black, other)
- Stroke Subtype (Infarction, primary intracerebral haemorrhage, subarachnoid haemorrhage or unknown/unclassified)
- Glasgow Coma Scale [GCS]
- BI at 7-10 days post stroke (0-9: severe disability, 10-14: moderate disability, 15-19: mild disability or 20: independent)

Although missing data were present in the SLSR in some of these variables, the overall amount of missing data in the variables discussed above was low. Rather than exclude participants with missing baseline data, reducing the number of cases available with complete outcome data, missing baseline variables were imputed using methods detailed below. The imputations were carried out on the entire SLSR cohort.

Data on age and sex were available for all patients. Ethnicity and subtype were both recorded on the register with an unknown category. Overall 2.5% of the SLSR have unknown ethnicity and 8.8% have an unknown or unclassified stroke subtype. For subtype this category was retained as these participants form a distinct group known to have had a stroke but for whom it is not possible to determine the exact subtype.

BI and GCS recorded after stroke were missing in 24.7% and 3.9% of patients respectively. Among those with missing BI 40.8% died within two weeks of the stroke. In analysis BI was treated as a 4 level ordinal variable using categories described in Chapter 4 Section 4.4.3.1. Ninety-six percent of all other patients dying within two weeks of stroke were categorised as 'severely disabled' using the Barthel. Therefore, missing BI data for those dying within 14 days were imputed as 'severely disabled'. This left 10.1% of the SLSR with unknown BI. A hotdeck approach (described in Chapter 2 Section 2.2.3.2) was employed to impute all unknown ethnicity and GCS's

and the remaining unknown BI scores by imputing scores using those from participants with similar characteristics. This resulted in a single dataset with no missing data in the baseline variables utilised in the simulation studies.

5.2.3.2 Complete data sample

In total 1604 SLSR participants had complete data until five years after stroke or death, which ever was soonest. Participants who died earlier, were eligible for less follow-ups than those surviving beyond 5 years. Consequently, among all participants with complete data, a greater proportion died sooner after stroke compared to the corresponding proportion among all SLSR participants. To obtain a dataset with complete outcome data in which the distribution of deaths was as similar as possible to the SLSR, a random sample of participants was extracted from all 1604 participants with complete data.

To achieve this, the proportion of all complete cases dying by each follow-up, which were required to produce a death distribution similar to the SLSR was estimated as illustrated in Table 5.1. The proportion of participants in the SLSR dying within each follow-up period is provided in the first column, and the number of participants with complete data prior to death in the second. The number of complete cases selected was maximised when all 319 patients surviving to at least five years after stroke were included. When these 319 participants were assumed to represent 44.7% of the sample, as in the SLSR, the estimated total sample size was 713.65. The expected number of complete cases dying in each follow-up period was then obtained by multiplying this total by the SLSR proportion. Finally, the proportion of complete cases required to achieve the desired distribution of deaths, and to define probability of inclusion, was calculated as $\frac{ExpectedNo.}{ObservedNo.}$.

For each participant, i, a random draw, $u_i \sim U(0,1)$ from a uniform distribution was made. A participant was then selected for inclusion in the dataset when

Table 5.1: Complete dataset - sample size

Time of death	SLSR proportion	No. of complete cases	Expected no.	Expected/Observed
0 - 3 months	0.281	885	200.54	0.227
3 months - 1 year	0.082	167	58.52	0.350
1 - 2 years	0.053	98	37.82	0.386
2 - 3 years	0.058	64	41.39	0.647
3 - 4 years	0.043	38	30.69	0.808
4 - 5 years	0.037	33	26.41	0.880
5+ years	0.447	319	319	1.000
Total			713.65	

$u_i < p(inclusion|D = d_{i,})$, where D is the period in which death occurred and $p(inclusion|D = d_{i,})$ is the probability of inclusion given death occurred in follow-up period, d.

5.2.3.3 Missing data scenarios

When presented with a dataset with missing data it is not possible to distinguish between missing data machanisms (See Chapter 2 Section 2.2.5.7). Therefore, it was not possible to simulate missing data using exactly the same mechanism behind missingness in the SLSR, as it is not known. Instead, four scenarios were derived in which different assumptions about the missing data were made. This was to allow comparisons of methods for handling missing data when specific groups of patients were more likely to drop out or miss a follow-up than the others; an assumption which is likely but untestable.

In all scenarios missing data were simulated to ensure that at each follow-up the rates of dropout, intermittent missingness and completed follow-ups reflect those observed in the SLSR (shown in Table 5.2). The first scenario assumes that all participants have an equal probability of having incomplete follow-ups and so are

MCAR. The second assumes that missing data are dependent only on baseline characteristics, which can be adjusted for in analyses, and so MAR. In addition to baseline characteristics, the third scenario allows the probability of missingness to be dependent on time of death, which, while known is not routinely adjusted for in analyses of prevalence rates prior to death. In the final scenario, follow-up BI, which in exploratory analyses was found to be the outcome most strongly associated with dropout, was used to simulate a mechanism in which participants who were more severely disabled were most likely to be missing. As the BI score on which missing data are dependent then becomes unknown, this scenario represents data MNAR.

In all four scenarios dropout was first introduced into the dataset in survivors at three months, at the rate outlined in Table 5.2, then at one year in those who had not dropped out at three months and subsequently at two, three, four and five year follow-ups. Next, intermittent missingness was introduced at each follow-up. For a follow-up to be intermittently missing it was required that the participant must be alive and not have completed at least one future follow-up otherwise the current follow-up would be considered the time of dropout. Therefore, intermittent missingness was simulated at a given time point in the subset of patients who were alive, and had not been identified as a dropout at the proceeding follow-up.

Table 5.2: Follow-up status of all SLSR patients (1995-2005) from three months to five years after stroke

n(%)	3 month	1 year	2 years	3 years	4 years	5 years
F/up complete	1595(50.7)	1472(46.8)	1293(41.1)	1155(36.7)	1040(33.1)	817(26.0)
Dropout	228(7.3)	183(5.8)	222(7.1)	217(6.9)	245(7.8)	318(10.1)
Missed f/up	437(13.9)	340(10.8)	318(10.1)	273(8.7)	228(8.7)	253(8.1)
Died	885(28.1)	1150(36.6)	1312(41.7)	1500(47.7)	1632(47.7)	1757(55.9)

For every participant, i, a series of random draws

$$u_{i,t,1} \sim U(0,1) \; and \; u_{i,t,2} \sim U(0,1)$$

where t =3m, 1, 2,3 ,4 ,5 year follow-up, were made. Eligible participants, i were then identified as a dropout at time T if $p_i(drop|T=t) < u_{i,t,1}$ and intermittently missing if $p_i(int|T=t) < u_{i,t,2}$. Individual probabilities of dropout and intermittent missingness were derived to reflect the missing data mechanisms assumed in each of the four scenarios. The methods used to derive these probabilities are described below.

5.2.3.4 Scenario I: Missing Completely At Random

To obtain data that was MCAR, the probabilities of dropout and intermittent missingness at each follow-up, were defined using the corresponding rates in all surviving SLSR participants (Table 5.2).

At three months after stroke, $p_i(drop|T = 3m)$ =proportion of SLSR survivors who dropped out at three months. Rates of dropout at one year and beyond are made up of a combination of participants who dropped out prior to the follow-up and new dropouts. As the proportion already dropped out at each follow-up varied from simulation to simulation, to ensure overall rates matched the fixed SLSR rates, the probability of new dropout was defined as

$$p_i(drop|T=t) = \frac{exp_t - obs_t}{N_t}$$

where

t=follow-up at 1, 2, 3, 4 or 5 years

exp_{SLSR_t}=expected dropouts at t = number survivors*SLSR dropout rate

obs_t=number of survivors at t who already dropped out

N_t=number of survivors at t.

The probability of intermittently missed follow-ups was then defined as $p_i(int|T = t)$ =rate of missed follow-ups in all SLSR participants at time t who were alive and not dropped out by $t + 1$.

5.2.3.5 Scenario II: Missing At Random

In the second scenario, missingness was assumed to depend on observed baseline data. To achieve this, a series of logistic regression models were fitted to the SLSR dataset. First models to predict dropout among survivors at each follow-up were fitted. Further models were then fitted to predict intermittently missed follow-ups among participants who were alive and who had not dropped out. Fitted models were then used to derive individual probabilities of dropout, $p_{fit_i}(drop|t)$, and intermittent missingness $p_{fit_i}(int|t)$ for all participants at each time point,t, based on baseline covariates.

Covariates included in the models were age, gender, ethnicity, stroke subtype, 7-10 day Barthel and GCS (stroke severity). The distribution of probabilities in the complete cases were very similar to those observed in the whole SLSR.

Fitted probabilities at time point t, were used to determine the $p_i(drop|t)$ and $p_i(int|t)$ used in simulations. To ensure that the overall levels of missing data matched the SLSR rates it was necessary to rescale the probabilities to ensure the average probability of dropout and intermittent missingness at each time point was equal to the proportion with missing data in the whole SLSR.

Starting at three months the probability of dropout was calculated as:

$$p_i(drop|T = 3m) = p_{fit_i}(drop|3m) * \frac{p(drop_{SLSR}|3m)}{mean(p_{fit_i}(drop|3m))}$$

where

$p(drop_{SLSR}|3m)$=Probability of dropout in SLSR at 3 months.

Participants were then marked as dropped out at the three month and all future follow-ups until death if $p_i(drop|T = 3m) < u_{i,3m,1}$. Next probability of dropout at one year, then two years and so on, was calculated by rescaling the fitted probabilities to ensure the mean dropout probability at each time point was the same as in the SLSR. The following formula was applied at follow-ups after three months:

$$p_i(drop|T = t) = p_{fit_i}(drop|t) * \frac{p(drop_{required}|t)}{mean(p_{fit_i}(drop|t))}$$

where

$p(drop_{required}|t) = \frac{exp_t - obs_t}{N_t}$

t=follow-up at 1, 2, 3, 4 or 5 years

exp_{SLSR_t}=expected dropouts at t = number survivors*SLSR dropout rate

obs_t=number of survivors at t who already dropped out

N_t=number of survivors at t.

After missing data due to dropout had been added to the datasets, intermittent missingness among those not dropped out was added. The probability of intermittent missingness used in the simulations was derived from the fitted probabilities from the SLSR and rescaled to ensure the mean probability in the non-dropouts at each time point was the same as in the SLSR. To achieve this, the following formula was applied:

$$p_i(int|T = t) = p_{fit_i}(int|t) * \frac{p(int_{required}|t)}{mean(p_{fit_i}(int|t))}$$

where

$p(int_{required}|t) = \frac{exp_t}{N_t}$

t=follow-up at 3 months, 1, 2, 3, 4 or 5 years

exp_{SLSR_t}=expected missing at t = number survivors*SLSR miss rate

N_t=number of survivors at t.

5.2.3.6 Scenario III: Missing At Random - dependent on time of death

In exploratory analyses described in the previous chapter, patients dying soonest had the poorest outcomes prior to death, particularly with regards to disability level. In the complete data sample, time of death during the five year follow-up period was known for all participants and as such was used to simulate a second MAR scenario (MAR(d)). As time of death occurs after the measurement of prevalence in survivors, it is not normally considered as a predictive factor or adjusted for when estimating prevalence of poor outcome, and so, although it was measured for all participants and used to simulate the missing data patterns, time of death was not explicitly adjusted for in the analyses of the simulated data.

Methods used in this scenario were identical to those applied in scenario II with the exception of covariates included in the logistic regression models. In this scenario, time of death (categorised as before 3 months, 3 months - 1 year, 1-2, 2-3, 3-4, 4-5 or more than 5 years after stroke) was included in addition to the baseline covariates.

5.2.3.7 Scenario IV: Missing Not At Random

Analyses presented in Chapter 4 suggested that lower Barthel scores are associated with earlier dropout as well as death. In the final scenario, missing data were assumed to be associated with current Barthel score. A linear relationship between Barthel score and probability of dropout and intermittent missingness was assumed.

In Table 5.3 the dropout rate at five years among participants who completed a four year follow-up is broken down by actual Barthel Score at four years. Participants who had a Barthel score in the three lowest possible scores were three to four times more likely to drop out before five years than those who had scores from 18-20. Similar trends were observed at other time points. As shown in Chapter 4, the average Barthel score dropped more rapidly between the two follow-ups prior to

dropout than between any other two time points, so it is possible that the Barthel scores of those who dropped out at five years would have been even lower than those observed at four years. Therefore in the simulation, to produce a MNAR mechanism that would allow for this, it was assumed that those with a Barthel score of zero at a given follow-up were six times more likely to drop out and miss that particular follow-up than those with a score of 20.

Intermittent missingness was not so obviously associated with Barthel scores. However, to allow for a possible but weaker MNAR association here it was assumed that those with a Barthel score of zero were twice as likely to drop out.

Missing data were simulated to ensure that the above relationships held true and that the overall rates of dropout and intermittent missingness were the same as those observed in the SLSR as a whole. Letting p_{drop} be the true or desired average probability of dropout at a desired time point, across the whole sample and let m =the relative increase in the probability of dropout(i.e. m=6) for those with Barthel score of zero relative to those with a score of 20. Then let $a = m/20$ be the corresponding increases in probability of dropout associated with a one unit increase in Barthel score. Then with n_{tot} =total sample size in which the missing data are to be simulated, the overall desired number of dropouts equals $n_{tot} * p_{tot}$. As probability of dropout was assumed to be linearly associated with Barthel there then exists a value p such that the

$$n_{20}p + (1+a)n_{19}p + (1+2a)n_{18}p + ... + (1+20a)n_0p = n_{tot}p_{tot}$$

where $n_0, ...n_{20}$ are the number of participants with Barthel scores 0,...,20. Rearranging the equation and solving for p gives

$$p = \frac{n_{tot}p_{tot}}{n_{tot} + (an_{19} + 2an_{18} + ... + 20an_0)}.$$

Table 5.3: Dropout rates at five years after stroke in participants surviving at least five years broken down by Barthel score at four years after stroke

Barthel Index	All N	Dropped out by five years N((%))
0	6	2(33.3)
1	9	4(44.4)
2	17	5(29.4)
3	10	1(10.0)
4	8	3(37.5)
5	10	6(60.0)
6	2	0(0.0)
7	14	3(21.4)
8	7	2(28.6)
9	11	2(18.2)
10	14	2(14.3)
11	21	5(23.8)
12	18	2(11.1)
13	20	2(10.0)
14	28	4(14.3)
15	46	7(15.2)
16	44	8(18.2)
17	73	12(16.4)
18	107	12(11.2)
19	125	18(14.4)
20	378	52(13.8)

For each participant their individual probability of dropout at that time point is then calculated with the probability of dropout for those with a Barthel score of 20, notated as $p(drop|BI = 20)$ equal to p. It then follows that $p(drop|BI = 19) = (1+a)p$, $p(drop|BI = 18) = (1 + 2a)p$,....,$p(drop|BI = 0) = (1 + 20a)p$.

Individual probabilities of intermittent missingness can be defined in the same way.

The methods described above in scenario IV for simulating the missing data based on individualised probabilities of response, while ensuring overall dropout and intermittent missingness rates reflect the SLSR were then applied.

Although the missingness was dependent only on Barthel score, the other outcomes considered are correlated with Barthel to varying degrees. As Barthel score was used to simulate the missing data it was considered very strongly correlated with the MNAR mechanism. Inactivity was measured using the Frenchay Activities Index which at one year after stroke is correlated with Barthel score with r=0.654 and r=0.648 at five years. Therefore the activity level data will also be missing not at random, and the strength of the association with the mechanism was considered strong. The HADS-D score was moderately correlated with Barthel (r=0.380 and r=0.352) and so depression considered to be moderately associated with the MNAR mechanism. HADS-A was only weakly associated with Barthel (r=0.185 and r=0.223) and considered weakly associated with the MNAR process.

5.2.4 Data analysis methods

The simulation study focused on the effect of missing data mechanisms on estimates of prevalence of poor outcome across four domains: disability level, inactivity, anxiety and depression. The performance of various methods for handling missing data were compared to examine the effect that choice of method has under different as-

sumptions about the missing data mechanism.

Prevalence rates at one and five years post stroke were calculated. Considering estimates at two times points allowed for the impact of a higher rate of death, resulting in lower sample sizes, as observed at five years compared to one year, to be assessed.

Methods for handling missing data were described in detail in Chapter 2. Below, the techniques used to apply these methods to the SLSR data are described. For each of the methods estimates of prevalence rate, associated standard error and sample size were stored. These parameters were required to derive the performance measures summarised in section 5.2.5. The true prevalence was calculated as the proportion, p, of survivors in the dataset before missing data were introduced, who had a poor outcome in each of the four domains at one and five years. For each of the methods, the estimated prevalence, $\hat{p}$ was the rate of poor outcome in survivors without missing data. Unless otherwise stated, the standard error was calculated as $se(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where n is the number of participants included in the analyses.

5.2.4.1 Complete case analysis

In complete case analyses, only survivors with complete data up to the point of analyses were included. In other words, to be included in the calculation of prevalence at one year, the three month follow-up must also have been completed. At five years, all follow-ups must have been completed to be eligible for inclusion.

5.2.4.2 Available case analysis

All participants who had complete data at one or five year follow-ups were included in the respective available case analyses, regardless of whether or not they had completed previous follow-ups.

5.2.4.3 Last observation carried forward

In last observation carried forward, the missing disability, inactivity, anxiety and depression data, were imputed at one year and five years in survivors using data from the last completed follow-up. In the analysis of outcomes at one, participants without a three month follow-up were excluded as no information was available to carry forward. Similarly any participants alive at five years but who dropped out before the three month follow-up, and so contributed no follow-up data at all, were excluded from analyses.

5.2.4.4 Inverse probability weighting

Logistic regression models are used to model the probability of being observed among survivors at one and five years after stroke (see section 2.2.2 for more details on IPW). The models adjusted for the factors used to simulate MAR data (i.e. age, gender, ethnicity, stroke subtype, GCS and Barthel index at 7-10 days after stroke). Fitted probabilities from these models were then used to weight the data using the formula weight= $1/prob(obs)$.

The fitted probabilities from the model of response in participants the SLSR as a whole were examined. There was no evidence that slight skewness in the continuous variables in the model resulted in large weights being assigned to a some subjects. As two continuous variables were included in the model (age, and GCS (treated as continuous, though strictly ordinal)), the assumption that the relationship between these variables and the logit of the outcome (i.e. missing or complete follow-up) is linear was assessed by plotting the lowess smoothed curve of age and GCS against the logit of the outcome. The continuous variables were therefore included without transformation. The goodness of fit of the model was also assessed using the Hosmer and Lemeshow test and there was no evidence of a lack of fit (p=0.423 and p=0.658, for the models applied at one and five years respectively).

The distribution of the weights in participants with complete data are shown in Figure 5.2

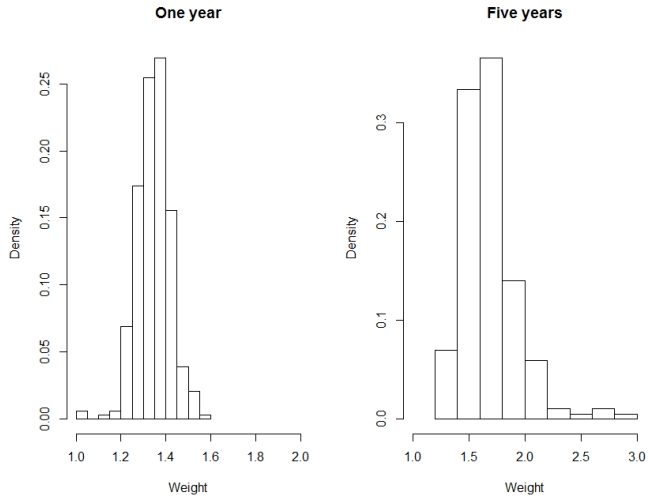


Figure 5.2: Distribution of inverse probability weights among those with complete data at one and five years after stroke

5.2.4.5 Mode imputation

Using the categorical outcome variables used to define poor outcome, missing data were imputed at one and five years using an age ic mode. The mode at one and five years was calculated in available data across four age groups, namely, <65, 65-74, 75-84 and $\geq$ 85 years.

5.2.4.6 Hotdeck imputation

A hotdeck approach was used to impute missing values by matching those with missing data to other subjects with similar baseline characteristics (age, gender, ethnicity, social class, stroke subtype and GCS).

The rrp.impute command in the 'rrp' package in R was used [204] to perform a nearest neighbour matching (See Section 2.2.3 for more details) with a single value imputed into each missing data point.

5.2.4.7 Regression imputation

Regression models were used to predict outcome at one and five years based on baseline information. Models were fitted to patients without missing information and then parameter estimates used to predict outcome for those with missing data. Models included terms for age, gender, ethnicity, stroke subtype, GCS, and BI at 7-10 days. The models imputed data using the categorical version of each variable. Logistic regression was used for the binary anxiety and depression data and proportional odds models used for the ordered three activity variable. Applying Brant test following a proportional odds regression showed no violation of the proportional odds assumption in the models for activity level. There was some suggestion that the effects of some variables may not be proportional in the models for disability level and so multinomial logistic regression was applied instead.

5.2.4.8 Multiple imputation

Multiple imputation by chained equations (see Chapter 2 Section 2.2.4 for detail) was carried out using the 'MICE ' package in R [114]. For each of the four outcomes, the imputations were performed independently. The outcome measures considered in this thesis are correlated to varying degrees, and so one outcome potentially predictive of another. However, imputing all four outcomes using a single imputation requires more complex (due to the inclusion of additional covariates) imputation models. The MICE procedure was the most computationally intensive step in the simulation study. Initial estimates of running time on a single PC for the simulation study was almost one calender year. This was reduced by upgrading hardware and by refining the R code used. Running the imputations seperately four times required less processing time than running once with all outcomes immputed simultaneously, and so independent imputations were one of the steps taken to minimise processing time required by MICE.

The imputation models for each of the outcomes included all the baseline variables

used to simulate the MAR data in scenario II, as summarised in Table 5.4. They also included the relevant outcomes at previous follow-up points.

Table 5.4: Format of predictor matirx used in MICE procediure

		Dependent Variables											
		Age	Sex	Ethnicity	GCS	Barthel	Subtype	Var3m	Var1	Var2	Var3	Var4	Var5
Independent Variables	Age	0	0	0	0	0	0	0	0	0	0	0	0
	Sex	0	0	0	0	0	0	0	0	0	0	0	0
	Ethnicity	0	0	0	0	0	0	0	0	0	0	0	0
	GCS	0	0	0	0	0	0	0	0	0	0	0	0
	Barthel	0	0	0	0	0	0	0	0	0	0	0	0
	Subtype	0	0	0	0	0	0	0	0	0	0	0	0
	Var3m	1	1	1	1	1	1	0	0	0	0	0	0
	Var1	1	1	1	1	1	1	1	0	0	0	0	0
	Var2	1	1	1	1	1	1	1	1	0	0	0	0
	Var3	1	1	1	1	1	1	1	1	1	0	0	0
	Var4	1	1	1	1	1	1	1	1	1	1	0	0
	Var5	1	1	1	1	1	1	1	1	1	1	1	0

Table summarises the variables included in the imputation models for outcomes from 3 months to five years after stroke. Baseline variables had no missing data and so were not imputed. A 1 indicated the variable was included in the model for the corresponding independent variable.
Abbreviations: GCS Glasgow coma score, BI Barthel Index.

where var3m,....var5 are the categorical outcome variables at 3m,.....5 years, respectively. At each follow-up time point only outcomes measured up to that point were included in the imputation models. For example at one year, only the outcome at three months was included in addition to the baseline variables. This approach was taken due to the substantial mortality rate among the participants.

All missing data were imputed in the same way regardless of whether it was miss-

ing due to dropout, was intermittently missing or missing due to death. After the imputations were carried out, those who had died were recoded as missing and so not included in prevalence estimates. An alternative approach would have been to define all participants who had died as having a poor outcome. However, this would likely lead to those who had died being far more likely to be identified as having a poor outcome than is true in practice, particularly when looking at rates of anxiety and depression. By five years, 55% had died, and so, had all future outcomes been used to impute data at the earliest time points, the majority of the data at, say, five years, would have itself have been imputed and for the substantial proportion who had died by five years, the information used in the imputation model would actually represent implausible values. However, the disadvantage of taking this approach is that for participants with intermittent missingness, potentially useful information from future follow-ups was not used.

As described in Chapter 2 Section 2.2.4.4, in MICE draws from the porterior distribution, which are used to multiply impute missing data, of the multivariate outcome are obtained by sampling from a series of conditional distribution estimated using appropriate regression models. To obtain values with which to impute the missing data, for anxiety and depression logistic regression models were used, proportional odds models for FAI and multinomial models for BI (via the logreg, polr and polyreg options in MICE).

Twenty imputations were carried out for each of the outcomes within each of the 1000 simulation datasets per scenario. Five iterations of the MICE cycle were used prior to saving each of the imputed datasets. While increasing the number of cycles helps to ensure that the random start values used in the imputations are not highly influential, and that the draws are independent of one another, processing time for MICE was again highly dependent on the number of iterations used. To explore the association between the imputed values and number of cycles, the mean and

standard deviation of imputed values across 50 iterations for a sample of five chains were examined and are provided in Tables B1 to B4 in Appendix B. Where the cycle has converged, random variation is expected, but there would be no evidence of a trend. Imputed values for all four outcomes did not show any evidence that the process did not converge quickly, and means and variation after five cycles were similar to that after further cycles.

The prevalence of poor outcomes at one and five years were then determined within each imputation dataset and the prevalence rates and standard errors combined using Rubin's rules.

5.2.4.9 Imputation for continuous scales

Imputation techniques were also applied to continuous forms of the scales used to define poor outcome. This was done to allow comparisons to be made between similar methods which impute continuous and categorical data.

After performing the imputations, new binary outcome variables were derived by dichotomising the variables using the same cut-off points as used to define poor/not poor outcome previously. Parameter estimates were then obtained using the same methods as described above.

Mean and median imputations were carried out using the same approach as for mode imputation above. Mean or median values in observed cases in four age groups were calculated using data from survivors without missing data and then used to impute values for other participants in the same age group with missing data.

Linear regression imputation again used a similar approach to regression imputation for categorical data, but rather than using logistic, linear regression models were applied.

Similarly, multiple imputation was carried out using the same process as for the binary outcomes, but with linear regression models used in the chained equations used to perform the imputations. As before five iterations were used over 20 imputations and figures summarising the mean and standard deviation of the imputed values are provided in Appendix B (Figures B5-B8). As before, similar variation was observed across all cycles, though for BI and FAI at some time points there was a downward trend prior to the means levelling out. However, this happened very quickly and by five iterations the mean was at a level at which it remained constant across further iterations.

5.2.5 Definition of performance measures

The ability of each of the above methods to produce unbiased and precise estimates of prevalence of poor outcomes was assessed by comparing estimates to the true values obtained from datasets without missing data. For each of the four scenarios, 1000 datasets with simulated missing data were created. The performance measures below were estimated by averaging across all 1000 datasets.

Bias: Average difference between prevalence, $\hat{p}$ estimates obtained in the simulated datasets and the true prevalence, p.

Precision: Average within simulation standard error of prevalence estimates.

5.3 Study 2: Effect of missing data on associations between baseline characteristics and outcome

5.3.1 Overview

The second study undertaken explored the impact of incomplete follow-up on associations between baseline characteristics and outcomes after stroke. Initially, it was planned that the simulation datasets produced in study one would also be used in the second study with a range of models applied to each dataset and parameter estimates from the models compared to one another and to the known 'true' value obtained by fitting the corresponding models to the complete dataset without missing data. However, many of the models described later in this section are complex and highly computationally intensive. This meant that it was not feasible to implement most of these models within a simulation study. Instead a series of models were applied to the full SLSR dataset, each of which made a different assumption about the underlying missing data mechanism. Although this meant that a known truth was not available, the range of models described below allow each made a different assumption about the underlying missing data mechanism meaning that the robustness of results to possible MNAR mechanisms could be assessed. The methods used in this study are described in the remainder of this chapter.

At the time of data extraction for this study, complete follow-up data and death records were available up to the end of 2012. Models were applied to data from follow-ups up to five years after stroke. Therefore, all SLSR participants who had a stroke between 1st January 1995 and 31st December 2007, and so eligible for five years of follow-ups, were included.

5.3.2 Model specification

The same four outcome measures as used in the simulation study described in the first half of this chapter were included. In the simulation study the focus was on estimating the prevalence of poor outcomes after stroke, with each measure dichotomised as poor or not poor. However, as described in chapter four, while anxiety and depression are most often dichotomised, activity level is often reported using a three level ordinal scale and disability level using a four point ordinal scale. In this study the categorical forms were used to explore the effect of incomplete data when models suitable for both binary and ordinal outcomes are applied.

The covariates included in the models were those previously described in section 4.4.2 and also used in the simulation study. These covariates provide a mix of binary, ordinal, nominal and continuous variables and have been shown to be associated with outcome, completeness of follow-up, or both (chapter 4). The covariates included were age at stroke onset, gender, ethnicity, stroke subtype, Glasgow coma score and Barthel index at 7-10 days after stroke.

Two of the outcomes, anxiety and depression, were binary and so models which were extensions of logistic regression were used to model these outcomes. For activity and disability level, both represented by an ordinal scale, exploratory analysis was conducted to determine the most appropriate models. The proportional odds model assumed that the effect of each covariate on the odds of being in one category relative to the one below is the same across all levels of the variable. As described in section 2.11, the Brant test can be used in a proportional odds model for non-correlated data. The test has not been extended to correlated data, therefore to give some indication regarding whether the proportional odds assumption is likely to hold in a longitudinal model, cross-sectional models were applied at each follow-up point. Where there was evidence that the assumption did not hold, the multinomial model was used instead.

As the estimates of outcome were made repeatedly over time, time since stroke needed to be included in the model. Previous analysis of post stroke disability suggest a sharp improvement up to three months to one year after stroke followed by a plateau and then gradual decline [158, 200, 201].Therefore the relationship between disability level and time since stroke is unlikely to be linear. Similarly, activity level is closely related to disability level and likely to have a similar relationship. Rates of anxiety and depression in the population of stroke survivors has been shown to remain stable over time, with new and relapsing cases equally likely at any time point up to 15 years after stroke [158, 162, 162]. The relationship between anxiety and depression and time is therefore less clear.

To determine the most appropriate form for the relationships between time and outcome a series of random effects models were applied. Logistic, proportional odds or multinomial models, as appropriate, with a random intercept, were fitted to the data. Each model included the baseline covariates listed above. In the first model time was included as a linear term only. Where significant a quadratic term was then added, followed by a cubic and so on until adding a higher power did not significantly improve the model fit. For two outcomes there was no evidence of a linear relationship and so quadratic and cubic models terms were added to ensure a non-linear relationship had not been missed. Where there was no evidence of a relationship a linear term was included only.

Non-linear mixed effects models, a class of models which allow for the non-linear association between covariates and outcomes, would potentially be useful in the analysis of the SLSR given the non-linear relationship between time and outcome described above. The advantage of such models is that they can often be specified such that all parameters have a meaningful interpretation [205]. The use of polynomial in equivalent linear models result in the inclusion of parameters which

are not easily interpreted and generally required more parameters. Further, while polynomials can offer good approximations of a non-linear relationship within the range of the observed data, when extending the model outside the observed data in general, a non-linear models result in more reliable predictions. However, non-linear models can be highly computationally intensive and require approximations of the maximum likelihood function to be made [205]. As some of the models used in this thesis are themselves difficult to fit and computationally intensive, and as the models used aimed to assess the relationship between baseline characteristics and outcome, rather than to describe the evolution of outcomes over time or predict outcomes out with the range of observed data, the use of simple polynomials as described above, were deemed sufficient.

Interactions between time and baseline covariates were not included in the models. The MNAR models applied to the data were highly computationally intensive and increasing the complexity of the models lead to increasing issued with convergence when fitting the models. In order to compare estimates across models, all models needed to include the same parameters, and therefor interactions were not included in any.

All covariates were considered individually in models that included only time in addition to the covariate of interest. Models were then applied which included and adjusted for all covariates simultaneously.

5.3.3 Marginal models

Four marginal models were fitted to the data; a GEE, two weighted GEEs and a multiple imputation GEE. Each of these models were previously described in Chapter 2, and their applications to the SLSR dataset are described below.

5.3.3.1 Generalised estimating equations

GEEs were applied to all four outcomes. For logistic and proportional odds models PROC GENMOD in SAS was used model the data [206] and the multGEE package in R used to fit multinomial GEEs [207]. Only the independent working correlation matrix is supported in analyses incorporating weights in SAS and in the multinomial GEEs in R. As only the logistic and proportional odds models could be fitted with different assumptions regarding the correlation structure, the effect of altering the assumptions was explored using logistic GEEs for the binary outcomes and proportional odds GEEs for the ordinal outcomes (even where there was violation of the proportional odds assumption). There was very little difference in parameter estimates from models with different working correlation assumptions and so an independence working correlation matrix was used in all models reported in the results chapter.

5.3.3.2 Weighted generalised estimating equations

As described in section 2.2.5, weighted GEEs (WGEEs) incorporate inverse probability weights and are appropriate for use when the missing data are monotone. The SLSR has both drop out and intermittent missingness, so in order to achieve a monotone missing data pattern, with missing data only due to exit from the study, intermittent missing values were imputed. In Chapter 4, exploratory analysis of the SLSR data did not reveal differences in outcomes according to pattern of intermittent missing data. Although a sudden temporary change in health status at the time of the missed follow-up cannot be ruled out, a MAR assumption appears plausible for the intermittently missing data. Further, in Chapter 6 the use of a simple LOCF was found to produce relative unbiased estimates of prevalence even in the presence of MNAR drop out in the SLSR. Therefore to impute the intermittently missing values a single imputation approach was used. As, by definition, intermittently missing values have at least one future observed value, there was assumed to be a linear relationship in the four outcomes between the next and last observed

outcome for each individual with values at the missed follow-ups imputed accordingly. Where the first follow-up was missed, the next observed follow-up was used to impute the data. Although this provided a monotone missing data pattern, the use of a single imputation methods for some of the missing data will result in artificially small standard errors and so estimates of variability and confidence intervals need to be treated with caution.

Weights were then constructed to estimate the probability of drop out at each follow-up point. Logistic regression models were then used to model the probability of being observed at a given follow-up in those who were still in the study at the end of the previous follow-up and these conditional probabilities multiplied to obtain the probability of drop out at any time point (see Chapter 2 Section 2.2.2.1 for details) The models included terms representing age, gender, ethnicity, stroke subtype, GCS and Barthel index at 7-10 days after stroke. Weights were then constructed as the inverse of the probability of being observed at each time point.

Two different WGEEs were applied, one representing a mortal analysis and the other an immortal analysis [69]. In the mortal WGEE, only participants alive at each follow-up point were included, by applying the logistic regression models and assigning IPWs to only those still alive. In contrast, in the immortal WGEEs the logistic regression models predicted the probability of being observed among all participants, including those who had already died who were classed as missing.

It was not possible to incorporate weights in a multinomial GEE using R, or any other standard statistical package, and so WGEEs were not applied to disability level post stroke, the outcome which violated the proportional odds assumption.

5.3.3.3 Multiple imputation and generalised estimating equations

Multiple imputation GEEs were fitted to the data. Data were imputed using PROC MI in SAS, using chained equations as described in section 2.2.4. The implementation of the multiple imputations was the same as in the simulation study (described in Section 5.2.4.8) with the exception of the imputation of disability level. PROC MI only allows for imputation of nominal variables using a discriminant function and so a multinomial logistic regression model was not used here.

Twenty imputations were used in each model.

5.3.4 Random effects models

Three models were applied which incorporated random effects and were fitted using maximum likelihood estimation. Before fitting the models exploratory analysis was carried out to explore the random effects structure. For each outcome an appropriate, i.e. a logistic, proportional odds or multinomial, GLMM with a random intercept was first applied to the data. A random slope was then added to allow for random variation in the rate of change over time and compared to the simpler random intercept model using a likelihood ratio test.

5.3.4.1 Generalised linear mixed models

Logistic, proportional odds and multinomial GLMMs were fitted to the data using PROC NLMIXED in SAS.

5.3.4.2 Shared parameter models

Shared parameter models were fitted to the binary and ordinal outcomes using PROC NLMIXED. As previously described in section 2.2.5.6 shared parameter models include a model for the outcome and a model for the drop out process which are linked via shared random effects. The outcomes of interest were modelled using mixed effects models which included a random intercept. In unadjusted analyses

single covariates were included along with time since stroke and all were added in the fully adjusted model. While death can occur at anytime and is therefore continuous in nature, dropout can only be defined as having occurred after a completed follow-up. If a participant dropped out after the four year follow up or died between 4 and 5 years after stroke, they were considered to have exited the study after the 4 year follow-up. Therefore time to exit from the study, for any reason, was treated as discrete and modelled using complimentary log-log models. Participants who missed a follow-up but later returned to the study were not deemed to have dropped out until after the latest follow-up in which they participated. The random effect from the outcome model was included as a covariate in the dropout model which also included age and gender. As multiple models were fitted, estimates of teh effect of age and gender on the drop out process have been omitted from the results tables in Chapter 7, as estimated coefficients were similar across all models. Age was highly significant with increasing likelihood of exit from the study with increasing age (beta ranged from 0.0201 (se=0.0024)) to 0.0247 (se=0.0026) and all $p<0.001$). Age was also included as exit from the study may be due to death or drop-out and gender is known to be associated with likelihood of death, but was not significant (beta ranged from -0.0400 (se=0.0554) to -0.0031 (se=0.0496), p=0.4669 to 0.9971)

It was not possible to fit models which would converge for disability level, which was modelled using a multinomial model nor for the adjusted proportional odds model for activity level. Additionally, the model for drop-out included only age and gender; and the complexity of the drop-out model increased there were again issues with model convergence.

5.3.4.3 Pattern mixture models

Pattern mixture models were also applied to the data. Participants were grouped according to the time at which they were last followed up. For example, a participant who died between the three month and one year follow-up would be classed as

having a final follow-up at three months. Similarly a participant who was still alive at five years but who completed only the three month follow-up would have had their final follow-up at three years after stroke. It was not possible to break drown the patterns further and separate those who dropped out and those who died due to sample size. Exploratory analyses of the SLSR presented in Chapter 4 showed that the outcome trajectories over time followed similar patterns in those who died and those who dropped out. Further, outcomes differed across all time points when grouped by time of death or dropout. It was therefore deemed more appropriate to define the patterns which incorporate time of exit rather than simply group by death or dropout. Participants who missed one or more follow-ups but completed at least one later follow-up were not deemed to have left the study at the time of the intermittently missed follow-up.

A series of dummy indicator variables were then derived to identify the point at which exit from the study occurred. These were then added to the random effects model described above in section 2.2.5.4 as interactions with the covariates in the model. This allowed the effect of the covariates to differ by time of exit from the study.

The parameters across the groups were also combined to give a single overall effect using the method described in 2.2.5.4. in which the estimated overall effect is obtained by weighting the model coefficients using the proportion of participants falling within each group. A pooled standard error was also calculated. However, a simple pooled standard error does not take into account the fact that the population proportions are also estimated. In section 2.2.5.4 the standard error for a simple case with two groups was described. No previous work has been reported in which standard errors have been derived for situations in which more than two groups are present. Further the main benefit and purpose here in fitting a pattern mixture model is the ability to compare parameter estimates across groups, rather than in

obtaining an overall effect size. Combined parameter estimates allow for comparisons with other models and simple pooled standard errors were used to calculate confidence intervals but it is acknowledged that the true standard errors are likely to be underestimated in the pooled results.

5.3.5 Comparing models

The results from the four GEEs all represent population averages effects and so are directly comparable with each other. Likewise, the random effects based models all represent subject specific effect and can be compared to each other. As described in section 2.1.2, it is possible to estimate marginal, or population average effects, from logistic regression GLMMs. The variance of the random effect was therefore used to rescale the parameter estimates and standard errors from the random effects based models. After being rescaled odds ratios and corresponding 95% confidence intervals were calculated and compared across all models.

In the absence of a more precise estimate of the relationship between population averaged effects and subject specific effects estimated using proportional odds and multinomial models, the same rescaling method was applied. This allowed for some comparison between the two model classes, but any differences need to be treated with caution.

All code used to fit the models above are available, with details provided in Appendix A.

5.4 Summary

In this chapter the methods used to explore the impact of incomplete follow-up data in the SLSR were described. The results of the simulation study are provided in Chapter 6 and the results from the study in which models were compared are

summarised in Chapter 7. Brief summaries are provided at the end of each of these chapters with a more detailed discussion given in Chapter 8.

Chapter 6

Results: Effect of missing data on prevalence estimates

6.1 Abstract

Background: Missing data in the South London Stroke Register are potentially *missing not at random* (MNAR). It is not known what impact the incomplete follow-up data have on estimates of the prevalence of outcomes after stroke or how best to handle the missing data.

Methods: A simulation study was carried out to compare estimates from complete and available case analyses (AC), inverse probability weighting (IPW), single imputation techniques and multiple imputation (MI). Missing data were simulated in a subset of the SLSR with complete data and reflected four different scenarios. Estimated prevalence of disability, inactivity, anxiety and depression, outcomes for which the strength of the association with probability of missingness differed in the simulated datasets, were compared to the true prevalence rate in the complete dataset.

Results: The lowest biases were observed following MI, followed by IPW. AC analysis was only marginally worse than MI and IPW and all three methods had similar standard errors. Single imputations resulted in substantial bias. When a strong

MNAR assumption was made AC analysis underestimated the true prevalence by up to 0.07 or 7% points for one outcome, and by up to 0.05 or <5% for the others. Using MI the maximum bias was 0.05 (5%).

Conclusions: Biases in the estimation of poor outcomes after stroke are not likely to be substantial when AC analyses are used. MI can reduce any bias, even when the data are MNAR.

6.2 Introduction

To determine the potential impact of missing follow-up data when estimating the prevalence of poor outcome after stroke, and to assess the performance of various missing data methods, a simulation study was conducted utilising data from the South London Stroke Register (SLSR). The methods used to carry out the study were described in detail in the Chapter 5 and the results presented here.

6.3 Comparison of SLSR and simulation datasets

6.3.1 Comparison of baseline characteristics

As previously described (Chapter 5), missing data were simulated in a subset of the SLSR in which participants had complete data. The missing data were simulated to reflect one of four scenarios in which the missing data mechanism varied. One thousand datasets were produced for each of the four scenarios. In the first scenario, the missingness reflected data that were MCAR, and so all participants had an equal chance of having missing follow-up data. In the second, the data were MAR, depending only on baseline participant characteristics and in the third, the missingness was MAR depending on time of death as well as baseline characteristics. In the final scenario, a MNAR mechanism which assumed missingness was dependent on current Barthel Index (i.e. level of disability) was used.

To ensure that participants included in the simulation datasets were representative of all SLSR participants baseline characteristics, rates of poor outcome and missing data patterns, averaged across each dataset within a scenario, were compared with the whole SLSR. The distribution of age, gender, ethnicity, stroke subtype and prior to stroke disability level are summarised in Table 6.1. Overall, participants included in the simulation datasets tended to be slightly older than those in the SLSR, more likely to be of white ethnicity and less likely to be black. They were also more likely to have had an ischaemic stroke. However, the size of the differences between the SLSR and the simulation datasets was relatively small.

Table 6.1: Comparison of demographic characteristics of the SLSR and simulation datasets

	SLSR	Sc I	Sc II	Sc III	Sc IV
		mean(sd)	mean(sd)	mean(sd)	mean(sd)
Total	3145	690(15.2)	692(14.9)	690(15.1)	689(15.1)
Age, mean	70.3	72.3(0.3)	72.3(0.3)	71.3(0.3)	71.4(0.3)
Gender,%					
Male	50.7	49.2(1.1)	47.2(1.1)	51.4(1.2)	50.7(1.1)
Ethnicity,%					
White	72.5	74.8(0.9)	75.4(0.9)	74.7(0.9)	74.5(0.9)
Black	19.4	17.0(0.8)	17.0(0.8)	17.8(0.8)	17.4(0.8)
Other	5.6	6.6(0.6)	5.5(0.6)	6.0(0.6)	6.4(0.6)
Unknown	2.5	1.4(0.4)	2.1(0.4)	1.5(0.4)	1.7(0.4)
Stroke Subtype,%					
Ischaemic	72.3	77.8(0.9)	74.6(1.0)	74.7(1.1)	73.1(1.1)
PICH	13.4	12.5(0.8)	13.1(0.8)	13.9(0.8)	13.9(0.8)
SAH	5.5	4.7(0.5)	5.9(0.5)	6.4(0.5)	8.1(0.6)
Other	8.8	5.1(0.7)	6.4(0.6)	4.9(0.7)	5.0(0.6)
7-10day Barthel,%					
Independent	58.6	56.7(1.0)	58.5(0.9)	58.5(0.9)	56.0(1.0)
Mild Disability	17.2	18.1(0.7)	15.8(0.6)	17.4(0.7)	17.8(0.7)
Moderate- Severe Disability	24.2	25.2(0.6)	25.7(0.6)	24.1(0.6)	26.2(0.6)

Table displays the mean (standard deviation) of proportions with a given characteristic averaged across 1000 simulated datasets in each scenario. Figures in the SLSR column are the means and proportions in the SLSR as a whole.

Scenario I assumes outcome data are MCAR, Scenario II assumes MAR, Scenario III assumes MAR with missingness dependent on time of death and Scenario IV assumes MNAR data with missingness dependent on current disability level

Abbreviations: SLSR South London Stroke Register, sd standard deviation, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage, MCAR missing completely at random, MAR missing at random, MNAR missing not at random.

6.3.2 Comparison of status at follow-up and rates of poor outcome

The average rates of poor outcome and follow-up status, for participants in the simulation datasets and the corresponding observed values from the SLSR, are presented in Table 6.2. The simulation studies were designed to mimic, as closely as possible, the missing data patterns observed in the SLSR. The average rates of completed follow-up, dropout, intermittently missed follow-ups and death are consistent across the datasets used in each of the four scenarios and very similar to the SLSR.

To compare the rates of poor outcome in the simulation datasets, rates were calculated in each dataset prior to the introduction of missing data. As described in Chapter 5 Section 5.2.3.3, to maximise the sample size of the complete dataset used in simulations, all participants alive at five years were included. Random samples of those who died earlier were then selected to ensure that the mortality rates in the simulated dataset mimicked those in the SLSR. Consequently, these is no variation in the numbers alive at five years as all participants were included in all datasets, prior to the introduction of missing data.

Overall, the prevalence rates in the simulation datasets were similar to those observed in the SLSR. At one year after stroke the average prevalence of depression in the simulation datasets was around 0.023 lower than the SLSR. Meanwhile, at five years the prevalence of anxiety was approximately 0.021 lower. For all other outcomes and time points, the average prevalence rates were within 0.005 of the corresponding SLSR rates.

Although the simulation datasets consisted of a sub-sample of the SLSR, selected as a result of the completeness of their data, the included participants appear to be representative of the whole SLSR cohort in terms of their baseline characteristics and rates of poor outcome.

Table 6.2: Comparison of rates of poor outcome after stroke in the SLSR and simulation datasets

	1 year					5 years				
	SLSR	Sc I	Sc II	Sc III	Sc IV	SLSR	Sc I	Sc II	Sc III	Sc IV
		mean(sd)	mean(sd)	mean(sd)	mean(sd)		mean(sd)	mean(sd)	mean(sd)	mean(sd)
Baseline sample size, n	3145	690(15.2)	692(14.9)	690(15.1)	689(15.1)	3145	690(15.2)	692(14.9)	690(15.1)	689(15.1)
Follow-up complete,%	46.8	46.7(1.5)	46.6(1.5)	46.8(1.4)	46.7(1.5)	26.0	26.3(1.1)	26.5(1.1)	26.4(1.0)	26.3(1.0)
Dropped out,%	5.8	6.5(0.7)	7.0(0.8)	6.6(0.7)	6.7(0.7)	10.1	10.2(0.6)	10.2(0.6)	10.2(0.6)	10.1(0.7)
Intermittent missingness,%	10.8	10.5(1.1)	10.8(1.1)	10.7(1.1)	10.6(1.1)	8.0	8.1(0.9)	7.8(0.9)	8.1(0.9)	8.2(0.9)
Died,%	36.6	36.3(1.3)	36.6(1.3)	36.6(1.4)	36.5(1.3)	55.9	55.4(1.4)	55.4(1.4)	55.3(1.4)	55.4(1.4)
Total alive, n	1994	450(6.8)	451(6.9)	450(6.8)	449(6.8)	1387	319	319	319	319
Moderate-Severe disability,%	25.5	25.2(1.1)	25.2(1.0)	25.2(1.0)	25.2 (1.1)	22.9	24.1	24.1	24.1	24.1
Inactivity,%	50.9	51.3(0.8)	51.3(0.7)	51.3(0.8)	51.3(0.8)	49.5	50.2	50.2	50.2	50.2
Anxiety,%	31.0	31.4(0.8)	31.5(0.8)	31.4(0.8)	31.4(0.8)	34.5	32.4	32.4	32.4	32.4
Depression,%	29.0	26.6(0.9)	26.6(0.8)	26.6 (0.8)	26.6(0.9)	29.5	29.4	29.4	29.4	29.4

Table displays the mean (standard deviation) of the proportions with a given outcome averaged across 1000 simulated datasets in each scenario. Figures in the SLSR column are the means and proportions in the SLSR as a whole.

Scenario I assumes outcome data are MCAR, Scenario II assumes MAR, Scenario III assumes MAR with missingness dependent on time of death and Scenario IV assumes MNAR data with missingness dependent on current disability level

Abbreviations: SLSR South London Stroke Register, MCAR missing completely at random, MAR missing at random, MNAR missing not at random.

6.4 Comparison of methods for handling missing data

As previously described in Section 5.2.5 two measures were used to assess the performance of a number of methods for handling missing data. For each method, within each scenario, the average bias (i.e. the difference between the estimated prevalence of poor outcome and the true prevalence) and precision (i.e. the average standard error of the prevalence estimate) were calculated. In the following sections the results for each of these measures are presented.

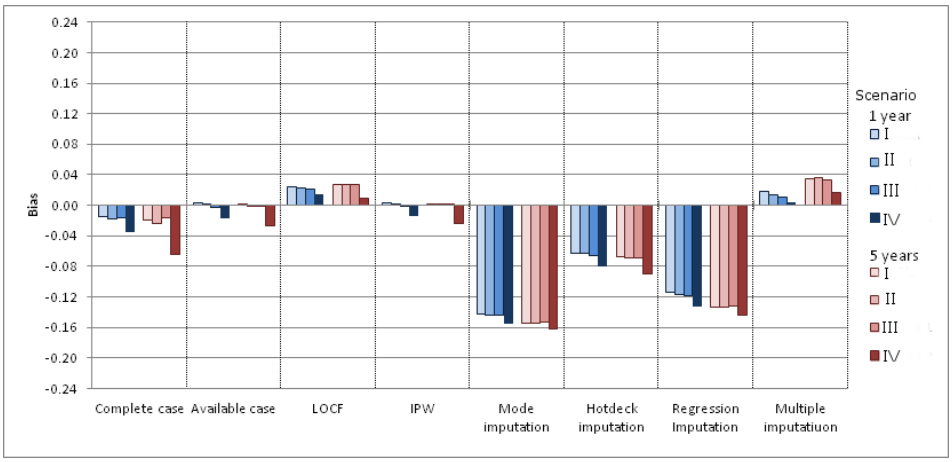
Four outcomes were considered in the simulation study. Two of these, anxiety and depression, both binary in nature, and moderately correlated with each other, produced similar results. For brevity the results for anxiety have been omitted from this chapter. Figures and tables for the anxiety outcome, corresponding to those presented within this chapter for the other outcomes, are available in Appendix C.

6.4.1 Depression

6.4.1.1 Bias

The bias associated with each of the methods applied to the binary form of the depression outcome is summarised in Figure 6.1 and Table 6.3. Using available case analyses, under the first three scenarios there was very little bias, with the largest difference between the estimated and true proportion with depression being 0.026 at five years. When only participants with complete data at all time points were included the bias was larger, with the magnitude of the differences between the true and estimated propotion being up to 0.023 in scenarios I to III and 0.035 and 0.064 points in scenario IV at one and five years, respectively.

Last observation carried forward resulted in an overestimation of the proportion with



Note: Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses with the following exceptions: complete case analysis included mean n=277 at one year and n=100 at five years; available case and IPW included mean=337 and n=206 and LOCF used data from mean n=419 and n=302 with data recorded at at least one previous follow-up.

Mean proportion with depression in complete data=0.27 at one year and 0.29 at five years.

Abbreviations: LOCF last observation carried forward, IPW inverse probability weighting

Figure 6.1: Bias of estimates of prevalence of depression using categorical missing data methods

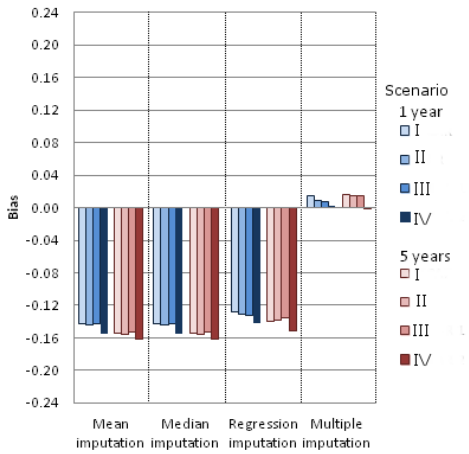
depression of up to 0.028 in scenarios I to III. The bias was lower in scenario IV at 0.015 at one year and 0.010 points at five years. Other single imputation methods resulted in substantial biases with the underestimation of up to 0.162 using mode imputation, 0.089 using hotdeck imputation and 0.144 using regression imputation.

IPW resulted in levels of bias which were very similar to those observed using complete cases only. Multiple imputation resulted in an overestimation of up to 0.017 at one year and 0.035 at five years, with the smallest biases observed in scenario IV.

The bias associated with imputation methods applied to the underlying scale before dichotomising the depression measure are summarised in Figure 6.2 and Table 6.3. The single imputation methods resulted in substantial underestimation of the true prevalence, at similar levels as when these methods were applied to the binary form of the depression measure.

When multiple imputation was applied to the continuous measure the bias was smallest in scenario IV. Across all four scenarios the bias was smaller when multiple imputation was carried out using the continuous form rather than the binary.

In addition to the bias associated with each methods the standard deviation of the prevalence estimates was also extracted from the simulations and is provided in Table D1 (Appendix D) for estimates of depression. At five years, all participants with complete data were included in all simulation dataset and so the standard deviation of the prevalence is equivalent to the standard error of the bias. However at one year, a random sample of those with complete one year data were included in each simulation and so there was some variation in the true prevalence rate used to calculate bias. The standard deviation of the prevalence is therefore not the same as the standard error of the bias, though offers an approximation. The largest standard



Note: Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses. Mean proportion with depression in complete data=0.27 at one year and 0.29 at five years.

Figure 6.2: Bias of estimates of prevalence of depression using continuous imputation methods

deviations were observed during complete case analysis (up to 0.079. Across the majority of simulations the standard deviations were less than 0.02. For a standard deviation of 0.08 a corresponding 95% confidence interval for the bias would have width +/-0.16 while for most scenarios and methods the confidence intervals would be at most +/-0.04.

Table 6.3: Bias associated with missing data methods when estimating prevalence of depression

		1 year				5 years			
		Sc I	Sc II	Sc III	Sc IV	Sc I	Sc II	Sc III	Sc IV
Categorical	Complete case	-0.015	-0.018	-0.016	-0.035	-0.019	-0.023	-0.017	-0.064
	Available case	0.003	0.000	-0.003	-0.017	0.001	-0.000	-0.002	-0.026
	LOCF	0.025	0.023	0.021	0.015	0.028	0.027	0.027	0.010
	IPW	0.002	0.002	-0.001	-0.015	0.001	0.002	0.003	-0.022
	Mode imputation	-0.142	-0.145	-0.144	-0.156	-0.155	-0.155	-0.153	-0.162
	Hotdeck imputation	-0.064	-0.063	-0.066	-0.080	-0.068	-0.069	-0.069	-0.089
	Regression imputation	-0.114	-0.117	-0.118	-0.133	-0.133	-0.134	-0.132	-0.144
	Multiple imputation	0.018	0.014	0.011	0.004	0.035	0.036	0.033	0.016
Continuous	Median imputation	-0.142	-0.145	-0.143	-0.156	-0.155	-0.155	-0.153	-0.162
	Median imputation	-0.142	-0.145	-0.143	-0.156	-0.155	-0.155	-0.153	-0.162
	Regression imputation	-0.128	-0.132	-0.133	-0.142	-0.140	-0.139	-0.136	-0.152
	Multiple imputation	0.014	0.010	0.007	0.002	0.016	0.014	0.015	-0.002

Scenario I assumes outcome data are MCAR, Scenario II Assumes MAR, Scenario III assumes MAR with missingness dependent on time of death and Scenario IV assumes MNAR data with missingness dependent on current disability level

Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses with the following exceptions: complete case analysis included mean n=277 at one year and n=100 at five years; available case and IPW included mean=337 and n=206 and LOCF used data from mean n=419 and n=302 with data recorded at at least one previous follow-up.

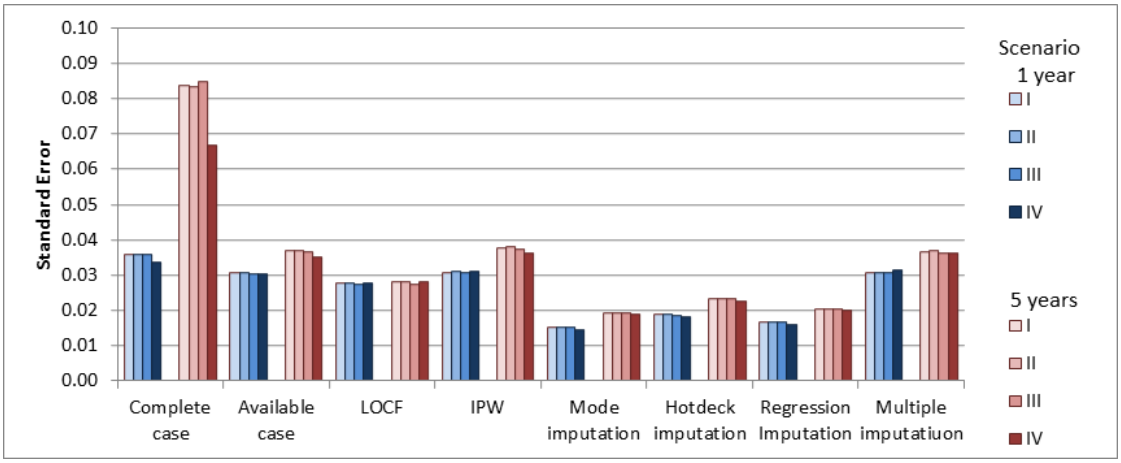
Mean proportion with depression in complete data=0.27 at one year and 0.29 at five years.

Abbreviations: LOCF last observation carried forward, IPW Inverse probability weighting, MCAR missing completely at random, MAR missing at random, MNAR missing not at random.

6.4.1.2 Precision

The average standard errors across the simulations from the methods applied to the binary form of the depression data are summarised in Figure 6.3 and Table 6.4. The largest standard errors were associated with complete case analysis at five years after stroke. Available case analysis, IPW and multiple imputation all produced standard errors that were similar to each other.

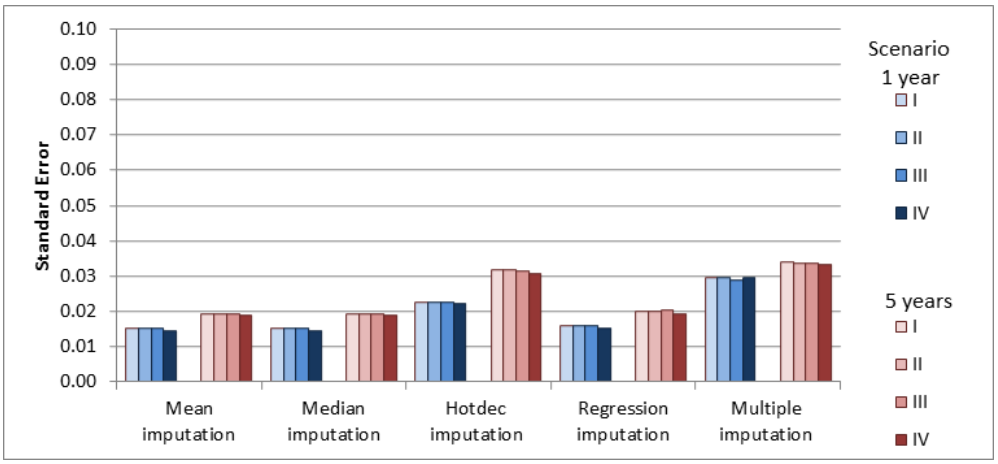
Among the single imputation methods, LOCF resulted in standard errors that were similar to available case, IPW and multiple imputation. Other single imputation methods, which were highly biased, had standard errors which were lower than any of the other methods.



Note: Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses with the following exceptions: complete case analysis included mean n=277 at one year and n=100 at five years; available case and IPW included mean=337 and n=206 and LOCF used data from mean n=419 and n=302 with data recorded at at least one previous follow-up.
Mean proportion with depression in complete data=0.27 at one year and 0.29 at five years.
Abbreviations: LOCF last observation carried forward, IPW inverse probability weighting

Figure 6.3: Standard error of estimates of prevalence of depression using categorical missing data methods

When imputations were carried out on the continuous depression scale, mean, median and regression imputation resulted in standard errors that were similar to those obtained from the corresponding imputations applied to the binary form (Figure 6.4 and Table 6.4).



Note: Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses. Mean proportion with depression in complete data=0.027 at one year and 0.029 at five years

Figure 6.4: Standard error of estimates of prevalence of depression using continuous imputation methods

Standard errors from multiple imputation on the continuous scale were slightly smaller than when multiple imputation was conducted using the binary form of the outcome measure.

Within each method the standard errors did not differ much across the four scenarios.

Table 6.4: Precision of methods when estimating prevalence of depression

		1 year				5 years			
		Sc I	Sc II	Sc III	Sc IV	Sc I	Sc II	Sc III	Sc IV
Categorical	True se	0.026	0.026	0.026	0.026	0.029	0.028	0.028	0.028
	Complete case	0.036	0.036	0.036	0.034	0.084	0.083	0.085	0.067
	Available case	0.031	0.031	0.030	0.030	0.037	0.037	0.037	0.035
	LOCF	0.028	0.028	0.027	0.028	0.028	0.028	0.027	0.028
	IPW	0.031	0.031	0.031	0.031	0.038	0.038	0.037	0.036
	Mode imputation	0.015	0.015	0.015	0.015	0.019	0.019	0.019	0.019
	Hotdeck imputation	0.019	0.0189	0.019	0.018	0.023	0.023	0.023	0.023
	Regression imputation	0.017	0.017	0.017	0.016	0.021	0.020	0.021	0.020
	Multiple imputation	0.031	0.031	0.031	0.031	0.037	0.037	0.036	0.036
Continuous	Median imputation	0.015	0.0153	0.015	0.015	0.019	0.019	0.019	0.019
	Median imputation	0.015	0.015	0.015	0.015	0.019	0.019	0.019	0.019
	Regression imputation	0.016	0.016	0.016	0.015	0.020	0.020	0.020	0.019
	Multiple imputation	0.029	0.030	0.029	0.030	0.034	0.034	0.034	0.033

Scenario I assumes outcome data are MCAR, Scenario II Assumes MAR, Scenario III assumes MAR with missingness dependent on time of death and Scenario IV assumes MNAR data with missingness dependent on current disability level

Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses with the following exceptions: complete case analysis included mean n=277 at one year and n=100 at five years; available case and IPW included mean=337 and n=206 and LOCF used data from mean n=419 and n=302 with data recorded at at least one previous follow-up.

Mean proportion with depression in complete data=0.27 at one year and 0.29 at five years.

Abbreviations: se standard error, LOCF last observation carried forward, IPW Inverse probability weighting, MCAR missing completely at random, MAR missing at random, MNAR missing not at random.

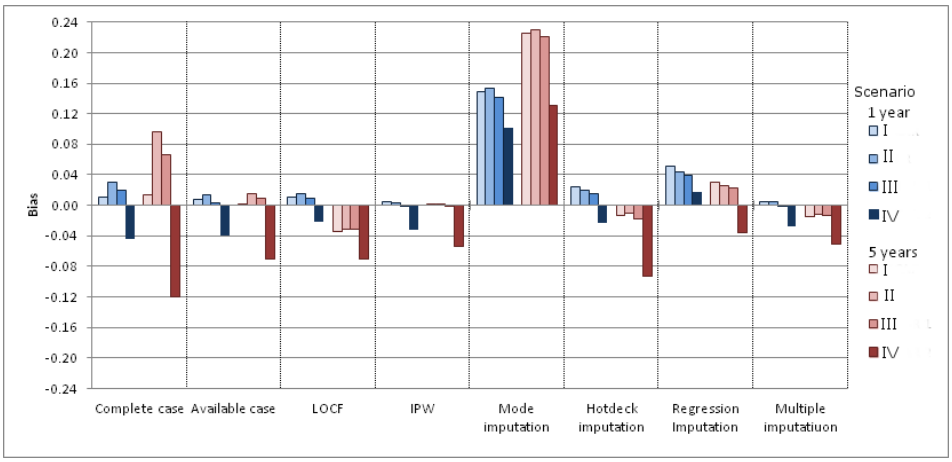
6.4.2 Inactivity

6.4.2.1 Bias

The bias associated with estimating rates of inactivity using methods applied to the categorical form of the inactivity measure are summarised in Figure 6.5 and Table 6.5. Estimates from complete case analyses were relatively unbiased in scenarios I to III, with the difference between the true and estimated proportion $\leq$0.016. However, in scenario IV, where the missing data were dependent on current level of disability, available case analysis led to an underestimation of 0.039 at one year and 0.071 at five years. Using only data from participants with complete data up to the time point being analysed in a complete case analysis resulted in overestimation of the prevalence in scenarios I to III, with the largest overestimation observed in scenario II at five years where it was 0.096. In scenario IV, the prevalence was underestimated by 0.044 at one year and 0.121 at five years after stroke.

Last observation carried forward lead to a small overestimation at one year in scenarios I to III, but an underestimation in scenario IV of 0.022. At five years the rate was underestimated in all four scenarios, with the largest difference between the true and estimated prevalence found in scenario IV. Among the other single imputation methods, mode imputation gave the most biased estimates, with the proportion who were inactive being overestimated by up to 0.230.

Both IPW and multiple imputation produced estimates that were less biased than available case analysis in scenario IV, where the data were MNAR. Across the other three scenarios they did not consistently produce estimates that were less biased than available case analysis.



Note: Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses with the following exceptions: complete case analysis included mean n=277 at one year and n=100 at five years; available case and IPW included mean=337 and n=206 and LOCF used data from mean n=419 and n=302 with data recorded at at least one previous follow-up.

Mean proportion who were inactive in complete data=0.510 at one year and 0.500 at five years.

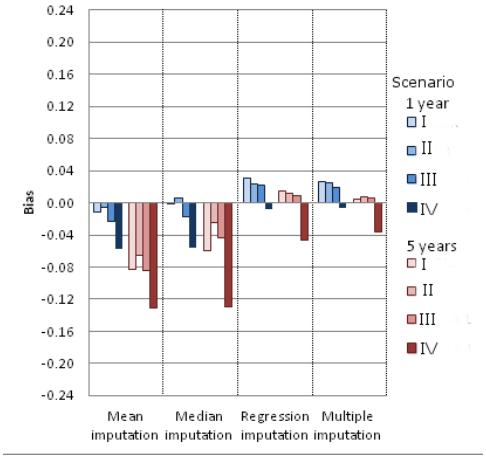
Abbreviations: LOCF last observation carried forward, IPW inverse probability weighting

Figure 6.5: Bias of estimates of prevalence of inactivity using categorical missing data methods

The biases associated with imputation methods applied to the continuous form of the activity data, before being dichotomised, are summarised in Figure 6.6 and Table 6.5. Mean and median imputations resulted in underestimation of the prevalence, while mode imputation using the categorical form resulted in an overestimation. The magnitude of the bias was less for mean and median imputation than for mode, but was still substantial, with the prevalence being underestimated by up to 0.130 at five years after stroke in scenario IV.

For both regression and multiple imputation, the bias was larger in some scenarios when the continuous measure was used, and in others, when the categorical form was used.

Estimates of the standard deviation of prevalence estimates of inactivity are provided in Table D2 of Appendix D. As with depression, the largest standard deviations were observed in complete case analysis, particularly when the data were MNAR (sd=0.0955). Standard deviations were also larger when the data were MNAR using all methods than under other mechanisms. In general the standard deviations when the data were not MNAR were up to 0.05. Therefore, 95% confidence intervals for the biases reported above would have width up to +/-0.2 when the data were MNAR and up to +/-0.1 under other scenarios.



Note: Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses. Mean proportion who were inactive in complete data=0.510 at one year and 0.500 at five years.

Figure 6.6: Bias of estimates of prevalence of inactivity using continuous imputation methods

Table 6.5: Bias associated with missing data methods when estimating prevalence of inactivity

		1 year				5 years			
		Sc I	Sc II	Sc III	Sc IV	Sc I	Sc II	Sc III	Sc IV
Categorical	Complete case	0.011	0.030	0.020	-0.044	0.014	0.096	0.066	-0.121
	Available case	0.008	0.013	0.003	-0.039	0.001	0.016	0.009	-0.071
	LOCF	0.011	0.015	0.010	-0.022	-0.034	-0.032	-0.032	-0.071
	IPW	0.006	0.002	-0.003	-0.036	0.002	-0.001	-0.001	-0.051
	Mode imputation	0.150	0.153	0.141	0.102	0.225	0.230	0.221	0.131
	Hotdeck imputation	0.024	0.0203	0.016	-0.023	-0.014	-0.010	-0.017	-0.093
	Regression imputation	0.051	0.044	0.039	0.017	0.030	0.025	0.023	-0.036
	Multiple imputation	0.005	0.004	-0.0010	-0.028	-0.014	-0.012	-0.013	-0.051
Continuous	Median imputation	-0.012	-0.006	-0.023	-0.058	-0.082	-0.066	-0.084	-0.131
	Median imputation	-0.000	0.006	-0.017	-0.056	-0.060	-0.024	-0.044	-0.130
	Regression imputation	0.031	0.024	0.022	-0.008	0.014	0.011	0.008	-0.047
	Multiple imputation	0.026	0.025	0.019	-0.007	0.005	0.008	0.007	-0.037

Scenario I assumes outcome data are MCAR, Scenario II assumes MAR, Scenario III assumes MAR with missingness dependent on time of death and Scenario IV assumes MNAR data with missingness dependent on current disability level

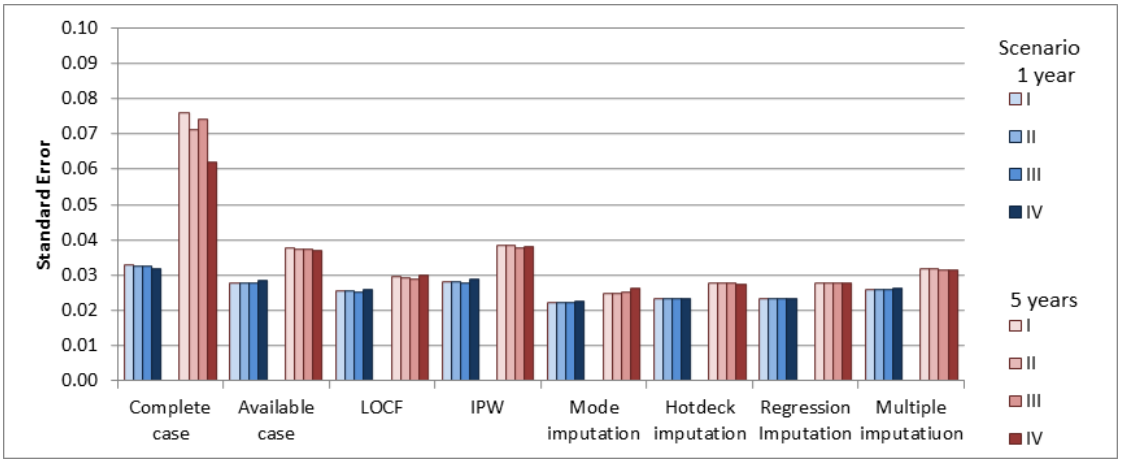
Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses with the following exceptions: complete case analysis included mean n=277 at one year and n=100 at five years; available case and IPW included mean=337 and n=206 and LOCF used data from mean n=419 and n=302 with data recorded at at least one previous follow-up.

Mean proportion who were inactive in complete data=0.510 at one year and 0.500 at five years.

Abbreviations: LOCF last observation carried forward, IPW Inverse probability weighting, MCAR missing completely at random, MAR missing at random, MNAR missing not at random.

6.4.2.2 Precision

The standard errors associated with methods applied to the categorical form of the activity data are summarised in Figure 6.7 and Table 6.6. As with depression, the largest standard errors were observed when complete case analysis was conducted at five years post-stroke. Overall the standard errors associated with the single imputation approaches, i.e. LOCF, mode, hotdeck and regression, were similar to each other and again lower than those from available case, IPW or multiple imputation.



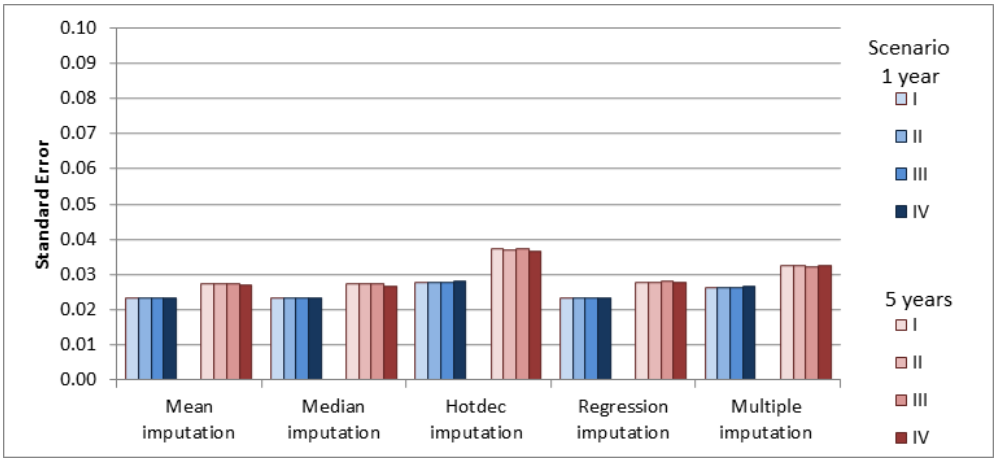
Note: Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses with the following exceptions: complete case analysis included mean n=277 at one year and n=100 at five years; available case and IPW included mean=337 and n=206 and LOCF used data from mean n=419 and n=302 with data recorded at at least one previous follow-up.
Mean proportion who were inactive in complete data=0.510 at one year and 0.500 at five years.
Abbreviations: LOCF last observation carried forward, IPW inverse probability weighting

Figure 6.7: Standard error of estimates of prevalence of inactivity using categorical missing data methods

The standard errors were smaller for multiple imputation than for the available case analysis or IPW.

When imputations were conducted using the continuous activity measure (Figure

6.8 and Table 6.6), the estimated standard errors were similar to the equivalent errors estimated following imputation of the categorical form. In multiple imputation, the standard errors were very slightly lower, and so closer to the true standard error, when the imputation was conducted using the categorical form.



Note: Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses. Mean proportion who were inactive in complete data=0.510 at one year and 0.500 at five years.

Figure 6.8: Standard error of estimates of prevalence of inactivity using continuous imputation methods

Table 6.6: Precision of methods when estimating prevalence of inactivity

		1 year				5 years			
		Sc I	Sc II	Sc III	Sc IV	Sc I	Sc II	Sc III	Sc IV
Categorical	True se	0.024	0.024	0.024	0.024	0.029	0.029	0.029	0.029
	Complete case	0.033	0.033	0.032	0.032	0.076	0.071	0.074	0.062
	Available case	0.028	0.028	0.028	0.029	0.038	0.038	0.038	0.037
	LOCF	0.025	0.026	0.025	0.026	0.029	0.029	0.029	0.030
	IPW	0.028	0.028	0.0278	0.028	0.038	0.038	0.038	0.038
	Mode imputation	0.0221	0.0221	0.0223	0.023	0.025	0.025	0.025	0.026
	Hotdeck imputation	0.023	0.023	0.023	0.023	0.028	0.028	0.028	0.027
	Regression imputation	0.023	0.023	0.023	0.023	0.028	0.028	0.028	0.028
	Multiple imputation	0.026	0.026	0.026	0.026	0.032	0.032	0.032	0.032
Continuous	Median imputation	0.023	0.023	0.023	0.023	0.027	0.028	0.028	0.027
	Median imputation	0.023	0.023	0.023	0.023	0.027	0.027	0.027	0.027
	Regression imputation	0.023	0.023	0.023	0.023	0.028	0.028	0.028	0.028
	Multiple imputation	0.026	0.026	0.026	0.027	0.033	0.033	0.032	0.033

Scenario I assumes outcome data are MCAR, Scenario II assumes MAR, Scenario III assumes MAR with missingness dependent on time of death and Scenario IV assumes MNAR data with missingness dependent on current disability level

Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses with the following exceptions: complete case analysis included mean n=277 at one year and n=100 at five years; available case and IPW included mean=337 and n=206 and LOCF used data from mean n=419 and n=302 with data recorded at at least one previous follow-up.

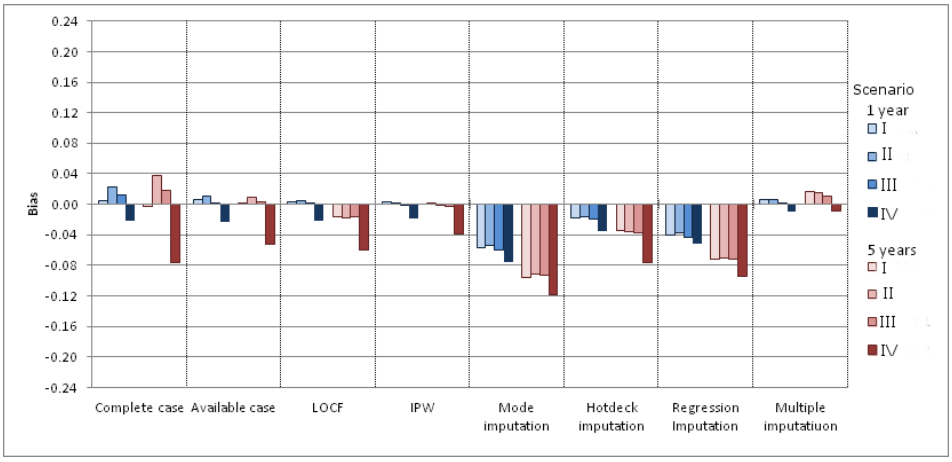
Mean proportion who were inactive in complete data=0.510 at one year and 0.500 at five years.

Abbreviations: se standard error, LOCF last observation carried forward, IPW Inverse probability weighting, MCAR missing completely at random, MAR missing at random, MNAR missing not at random.

6.4.3 Disability

6.4.3.1 Bias

The biases associated with methods applied to the categorical disability measure are summarised in Figure 6.9 and Table 6.7. All methods underestimated the prevalence of moderate-severe disability in scenario IV, where missing data were dependent on current level of disability. The largest biases were associated with complete case analysis, mode imputation and regression imputation where the proportion with moderate to severe disability was underestimated by up to 0.119 at five years after stroke in scenario IV.



Note: Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses with the following exceptions: complete case analysis included mean n=277 at one year and n=100 at five years; available case and IPW included mean=337 and n=206 and LOCF used data from mean n=419 and n=302 with data recorded at at least one previous follow-up.
Mean proportion with moderate-severe disability in complete data=0.240 at one year and 0.250 at five years.
Abbreviations: LOCF last observation carried forward, IPW inverse probability weighting

Figure 6.9: Bias of estimates of prevalence of disability using categorical missing data methods

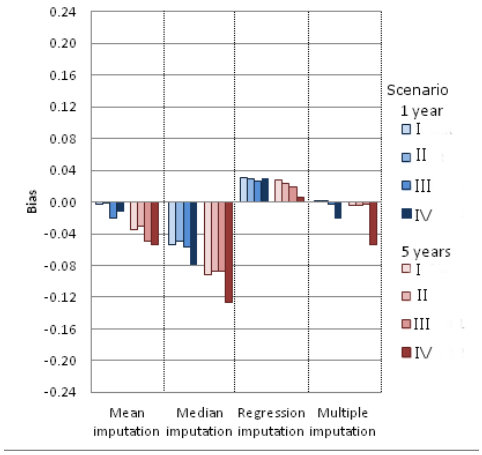
In scenarios I to III, there was very little bias observed in the available case analysis, though the prevalence was underestimated by 0.023 and 0.053 at one and five years after stroke respectively in scenario IV.

IPW produced the least biased estimates across scenarios I to III, but in scenario IV the prevalence was underestimated by 0.018 at one year and 0.039 at five years respectively. In scenario IV, the bias was smallest following multiple imputation where the true prevalence was underestimated by just 0.010 at both one and five years.

When imputations were applied to the continuous disability measure, mean and median imputation all led to underestimation of the true prevalence, with the largest biases observed in scenario IV (Figure 6.10 and Table 6.7). Regression imputation led to an overestimation.

Multiple imputation of the continuous scale produced unbiased estimates of prevalence in scenarios I to III, where the bias was lower than that observed when imputations were on the categorical form. However, the bias was greater in scenario IV using the continuous form.

Standard deviations of estimates of the prevalence are provided in Table D3 (Appendix D). The standard deviation of the estimates were up to 0.05, but for the majority of scenarios were less than 0.02, implying that 95% confidence intervals for the estimates of bias above would be at most +/-0.1 in width, and in many cases +/-0.04 wide.



Note: Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses. Mean proportion with moderate-severe disability in complete data=0.240 at one year and 0.250 at five years.

Figure 6.10: Bias of estimates of prevalence of disability using continuous imputation methods

Table 6.7: Bias associated with missing data methods when estimating prevalence of moderate-severe disability

		1 year				5 years			
		Sc I	Sc II	Sc III	Sc IV	Sc I	Sc II	Sc III	Sc IV
Categorical	Complete case	0.004	0.023	0.012	-0.022	-0.003	0.037	0.019	-0.077
	Available case	0.006	0.011	0.002	-0.023	0.001	0.009	0.004	-0.053
	LOCF	0.003	0.005	0.002	-0.022	-0.016	-0.018	-0.016	-0.060
	IPW	0.002	0.003	-0.002	-0.022	0.001	0.001	-0.003	-0.036
	Mode imputation	-0.058	-0.054	-0.060	-0.075	-0.097	-0.091	-0.093	-0.119
	Hotdeck imputation	-0.017	-0.017	-0.020	-0.035	-0.034	-0.036	-0.038	-0.076
	Regression Imputation	-0.042	-0.038	-0.043	-0.052	-0.072	-0.071	-0.071	-0.094
	Multiple imputation	0.006	0.006	0.002	-0.010	0.017	0.015	0.011	-0.010
Continuous	Mean imputation	-0.003	-0.001	-0.020	-0.013	-0.035	-0.031	-0.050	-0.054
	Median imputation	-0.054	-0.050	-0.056	-0.079	-0.092	-0.087	-0.088	-0.127
	Regression imputation	0.031	0.030	0.027	0.030	0.027	0.024	0.019	0.005
	Multiple imputation	0.001	0.001	-0.002	-0.022	-0.004	-0.005	-0.004	-0.054

Scenario I assumes outcome data are MCAR, Scenario II assumes MAR, Scenario III assumes MAR with missingness dependent on time of death and Scenario IV assumes MNAR data with missingness dependent on current disability level

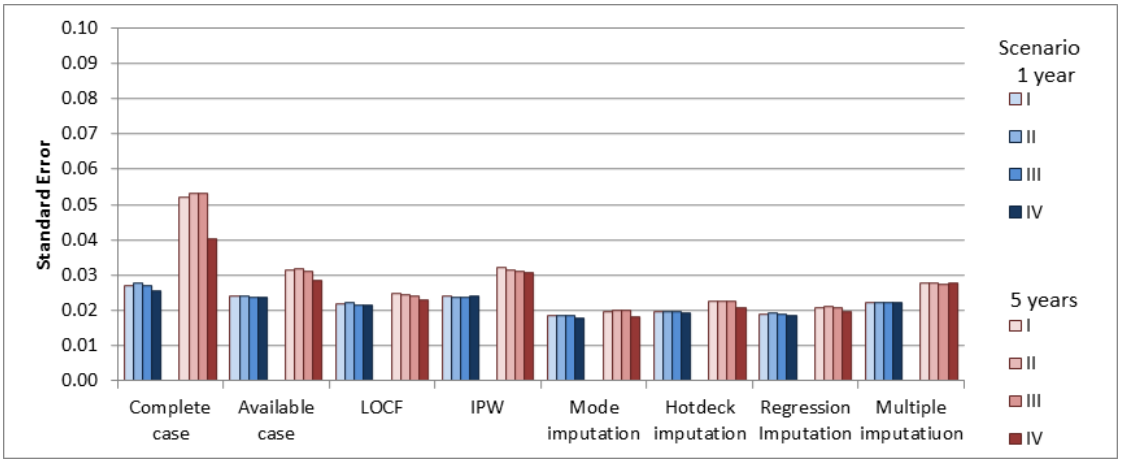
Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses with the following exceptions: complete case analysis included mean n=277 at one year and n=100 at five years; available case and IPW included mean=337 and n=206 and LOCF used data from mean n=419 and n=302 with data recorded at at least one previous follow-up.

Mean proportion with moderate-severe disability in complete data=0.240 at one year and 0.250 at five years.

Abbreviations: LOCF last observation carried forward, IPW Inverse probability weighting, MCAR missing completely at random, MAR missing at random, MNAR missing not at random.

6.4.3.2 Precision

The standard errors associated with methods applied to the categorical disability measure were largest for complete case analysis, followed by available case and IPW at five years after stroke (Figure 6.11 and Table 6.8). The standard errors following multiple imputation were lower than those from IPW or available case analysis. Standard errors from the single imputation methods were again lower than for the other methods.

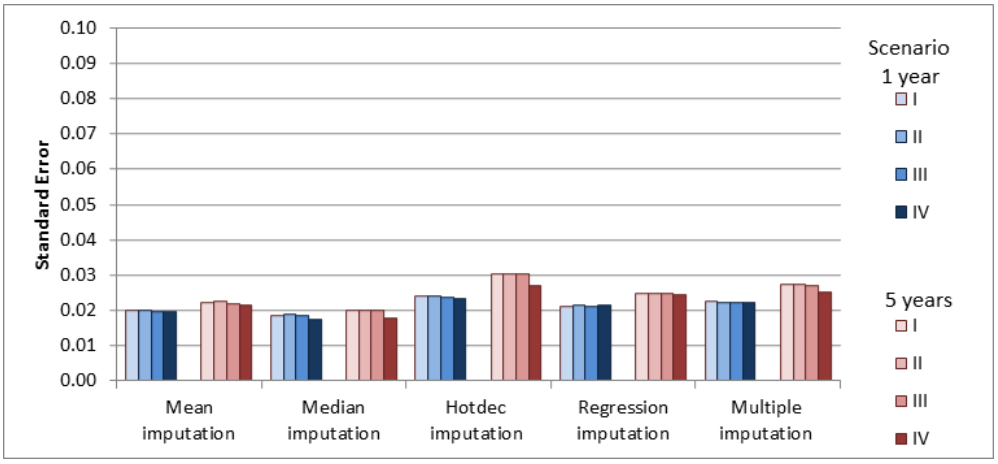


Note: Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses with the following exceptions: complete case analysis included mean n=277 at one year and n=100 at five years; available case and IPW included mean=337 and n=206 and LOCF used data from mean n=419 and n=302 with data recorded at at least one previous follow-up.
Mean proportion with moderate-severe disability in complete data=0.240 at one year and 0.250 at five years.
Abbreviations: LOCF last observation carried forward, IPW inverse probability weighting

Figure 6.11: Standard error of estimates of prevalence of disability using categorical missing data methods

The standard errors obtained from single imputation methods applied to the continuous form of the data were similar to those from the corresponding methods applied to the categorical disability measure (Figure 6.12 and Table 6.8). The standard errors from multiple imputation were also very similar to those obtained when the

imputations were performed on the categorical measure.



Note: Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses. Mean proportion with moderate-severe disability in complete data=0.240 at one year and 0.250 at five years.

Figure 6.12: Standard error of estimates of prevalence of disability using continuous imputation methods

Table 6.8: Precision of methods when estimating prevalence of moderate-severe disability

		1 year				5 years			
		Sc I	Sc II	Sc III	Sc IV	Sc I	Sc II	Sc III	Sc IV
Categorical	True se	0.020	0.020	0.020	0.020	0.024	0.024	0.024	0.024
	Complete case	0.027	0.028	0.027	0.026	0.052	0.053	0.053	0.040
	Available case	0.024	0.024	0.024	0.024	0.032	0.0318	0.031	0.028
	LOCF	0.022	0.022	0.022	0.022	0.025	0.025	0.024	0.023
	IPW	0.024	0.024	0.024	0.024	0.032	0.032	0.031	0.031
	Mode imputation	0.019	0.019	0.018	0.018	0.020	0.020	0.020	0.018
	Hotdeck imputation	0.020	0.020	0.020	0.019	0.023	0.023	0.022	0.021
	Regression imputation	0.019	0.019	0.019	0.019	0.021	0.021	0.021	0.020
	Multiple imputatiuon	0.022	0.022	0.022	0.022	0.028	0.028	0.027	0.028
Continuous	Median imputation	0.020	0.020	0.020	0.020	0.022	0.023	0.022	0.022
	Median imputation	0.019	0.019	0.019	0.018	0.020	0.020	0.020	0.018
	Regression imputation	0.021	0.021	0.021	0.021	0.025	0.025	0.025	0.024
	Multiple imputation	0.022	0.022	0.022	0.022	0.027	0.027	0.027	0.025

Scenario I assumes outcome data are MCAR, Scenario II assumes MAR, Scenario III assumes MAR with missingness dependent on time of death and Scenario IV assumes MNAR data with missingness dependent on current disability level

Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses with the following exceptions: complete case analysis included mean n=277 at one year and n=100 at five years; available case and IPW included mean=337 and n=206 and LOCF used data from mean n=419 and n=302 with data recorded at at least one previous follow-up.

Mean proportion with moderate-severe disability in complete data=0.240 at one year and 0.250 at five years.

Abbreviations: se standard error, LOCF last observation carried forward, IPW Inverse probability weighting, MCAR missing completely at random, MAR missing at random, MNAR missing not at random.

6.5 Summary and conclusions

The results from this simulation study highlight the potential impact incomplete follow-up data may have on estimates of the prevalence of poor outcome after stroke among participants in the SLSR as well as comparing the ability of various missing data methods to correct for this bias.

Exploratory analysis of the SLSR suggested that missingness, particularly dropout rather than intermittent missingness, was not missing completely at random. As described in Chapter 4 some groups of participants (for example, younger stroke survivors) were more likely to drop out than others, Further, declines in health across the domains considered in this simulation study were observed prior to dropout and death, with participants most likely to drop out in the years leading up to death.

The third scenario considered in this simulation study was designed to reflect the patterns observed in the SLSR as closely as possible, with a MAR pattern introduced into the simulation datasets by assuming missingness was dependent on both baseline characteristics and time of death (MAR(d)). The first two scenarios made less strict assumptions by assuming the data were MCAR or MAR dependent on baseline characteristics only. The final scenario made the strictest assumptions by creating missingness patterns in which missingness was dependent on currently level of disability, an assumption that is plausible but untestable in the SLSR dataset.

When estimates of the prevalence of poor outcome were obtained from the simulation datasets after applying various methods for handling missing data the largest biases were observed when data were assumed to be MNAR depending on current disability level. There were few differences observed in the levels of bias associated with missingness that was MCAR, MAR or MAR(d).

There was wide variation in the performance of methods for handling missing data.

Overall, multiple imputation provided relatively unbiased estimates of prevalence. Even when data were MNAR, dependent on disability level, the associated bias was minimal. Available case analysis and estimates obtained using IPW were associated with similar levels of bias to that resulting from multiple imputation, although the standard errors were larger.

The maximum bias associated with these three methods, as well as LOCF, which also produced largely unbiased extimates, across are summarised in Table 6.9.

Table 6.9: Summary of bias in estimates of poor outcome after stroke across under MNAR assumption

		Maximum absolute bias			
Outcome	True Prevalence	Available case	IPW	LOCF	MI
One year after stroke (% survivors with incomplete data=38.6%)					
Depression	0.290	-0.017	-0.014	0.015	0.004
Inactivity	0.509	-0.039	-0.033	-0.022	-0.028
Disability	0.255	-0.023	-0.018	-0.022	-0.010
Five years after stroke (% survivors with incomplete data=41.0%)					
Depression	0.295	-0.026	-0.025	0.010	0.016
Inactivity	0.495	-0.071	-0.054	-0.071	-0.051
Disability	0.229	-0.053	-0.039	-0.060	-0.010

Abbreviations: LOCF last observation carried forward, MI multiple imputation, MNAR missing not at random.

Under a MAR or MNAR mechanism, available case analysis would be expected to be biased, and MI and IPW biased under a MNAR mechanism. In this study the overall bias associated with each method was relatively low, even when the data were MNAR. The MNAR mechanism used in this simulation study resulted in increasing likelihood of being missing from a follow-up as level of disability at that follow-up increased. While assumptions were made that the level of disability was strongly predictive of missingness, the distribution of barthel scores, the measure of disability used, is highly skewed, with the majority of patients having only mild

or no disability. For example, at five years after stroke 39% of SLSR participants have no disability, achieving the maximum possible barthel score. Conversely, only 1% score zero, the lowest possible score and 10% score between zero and nine, the range classified as severe disability. Therefore, while drop out may be strongly associated with barthel score at a given time point and those who are most unwell are most likely to be missing, the actual impact on parameter estimates using observed follow-up data is minimal due to the fact that those who are most unwell represent only a small proportion of all stroke survivors.

In the simulation study IPW was used to weight responses with weights derived from a missingness model which included only baseline characteristics. On the other hand, the multiple imputation procedure used made use of data from previous follow-ups in addition to the baseline characteristics in the imputation model. Where the data were MNAR, bias was lower using MI than IPW or available case analysis. Although MI cannot remove all bias in this MNAR scenario, the inclusion of follow-up data, which will, to a certain extent be correlated with the missing values, may help to reduce biases arising from a potential MNAR mechanism in analyses of SLSR follow-up data.

The results also confirm the limitations and highlight the dangers of applying single imputation techniques. Using mean, median or mode imputation all resulted in large biases with the estimates of poor outcome being under or overestimated by as much as 0.200 or 20%. This demonstrates the dangers of applying these simple imputation techniques, which performed far worse than simply ignoring the missing data and analysing using available data only. While mean, median and mode imputations involved imputing age specific averages, hotdeck and regression imputation methods, also single imputation techniques, made use of more of the participants' available data to impute the missing values. These methods improved upon imputing averages but were still associated with large biases.

Last observation carried forward had the least biased estimates among these single imputation methods and was associated with a low level of bias overall and in fact produced estimates with similar, or lower, bias than available case analysis when the data were MNAR. Due to the nature of recovery after stroke, while there is often an initial period of recover, disability and activity levels tend to deteriorate after this initial recover period. While there is some fluctuation in scores, in general it seems reasonable that the category in which participants are classed at one follow up may be the same at the next.

Imputation methods were applied to both categorical outcomes and the original continuous scales from which the categories were derived. Overall there were few differences between the estimates from the categorical outcomes and continuous scales. In particular multiple imputation produced similar levels of bias in both cases.

In terms of the precision of the estimates of prevalence, the low standard errors associated with the single imputation methods, when combined with the highly biased estimates again highlight the potential dangers of such methods. Ninety-five percent confidence intervals for the prevalence estimate may not include the true prevalence and there is an increased likelihood of type I error when applying hypothesis tests.

Available case analysis resulted in standard errors which were similar to IPW, while Multiple imputation standard errors were marginally lower than available case. In general incorporating weights results in larger standard errors than those obtained in available case analysis. However where there is low variability in the weights used, then estimates will be closer to those observed for complete cases [33]. As shown in Chapter 5 Section 5.2.4.4, the weights used in the simulations, particularly

at one year were not highly variable, which may explain the similarity between the available case and IPW results.

Using multiple imputation, all cases are used in the analysis which in general increases precision and where the data are MAR and the correct imputation model has been used, it is expected that standard errors would be similar to those observed using the complete data, prior to the simulation of the missing data. In some cases the MI standard errors were the same as available case, however, where this occurred there were slight differences in levels of bias between the methods and as the standard error is a function of the prevalence estimate they are not directly comparable when the parameter estimates differ. Differences in bias were small so this may have minimal impact on the standard errors. Otherwise the standard errors from MI were marginally lower than available case when the imputation was performed on the categorical version of the outcome variable. For disability the standard errors were similar when multiple imputation was performed on the original barthel index and the derived categorical variable. However for the other two outcomes, standard errors were lower when the continuous version was used.

When estimating rates and proportions of outcomes at a given time point after stroke it would be recommended that MI, be used in addition to any available case analyses, particularly where the variable of interest is likely to be strongly related to dropout. However, standard errors associated with MI were larger than expected and so further work is needed to explore the imputation models used and determine the most appropriate set of predictors to include in the models as well as the type of model used to avoid potentially mis-specified or over-specified models which may lead to increased variation between imputation datasets and large standard error estimates

Chapter 7

Results: Effect of missing data on predictors of poor outcomes after stroke

7.1 Abstract

Background: In populations of people sharing a common condition identifying those at greatest risk of poor outcome is important. The South London Stroke Register (SLSR) collects data from stroke survivors which have often been used to estimate associations between characteristics at time of stroke and outcomes after stroke. Around one third of participants do not participate and this data may be *missing not at random* (MNAR) and the impact of the missing data on estimated associations is not clear.

Methods: Seven models (namely, generalised estimating equations (GEE), two weighted GEEs, multiple imputation GEE, generalised linear mixed models, pattern mixture and shared parameter models) which differ in the way they treat and make assumptions about the underlying missing data mechanism were applied to the SLSR data and parameter estimates compared. Two binary outcomes were analysed using logistic based models and proportional odds and multinomial models applied to two

ordinal outcomes.

Results: Parameter estimates were consistent across all the models, even those which allowed for a MNAR dropout process.

Conclusions: There was no evidence that the incomplete follow-up data in the SLSR biases estimates of associations between baseline characteristics and outcome after stroke.

7.2 Introduction

In this chapter the results of analyses conducted to explore the effect of missing data on associations between baseline covariates and outcomes in the South London Stroke Register (SLSR) are presented. A number of models, each making different assumptions about the missing data mechanism, were applied and the parameter estimates from the models were compared. The methods and models used were described in detail in Chapter 6 and the models and underlying assumption regarding the missing data mechanism are summarised in Table 7.1.

The first model was a standard GEE, followed by two weighted GEEs. In the first, the 'mortal' GEE, logistic regression models were used to derive weights in those who were alive at that time point, with no weights assigned to those who had died. In the 'immortal' GEE, logistic regressions compared those who were followed to those who had missing data, for any reason, including death. In models used to derive the weights included baseline characteristics, and so make the assumption that the drop-out process is explained by the characteristics of participants at the time of stroke. In the final model, multiple imputation was combined with the GEE (MIGEE). The MIGEE also assumes the data are MAR, but the imputation models include follow-up data observed at previous time points. The MIGEE therefore is assuming that the missing values themselves depend on baseline characteristics and previously observed values of the same outcome.

Table 7.1: Models and assumed missing data mechanism

Model	Mechanism
Marginal models	
GEE	MCAR
WGEE (Mortal)	MAR
WGEE (Immortal)	MAR
MIGEE	MAR
Random-effects models	
GLMM	MAR
Shared parameter model	MNAR
Pattern mixture model	MNAR

Abbreviations: GEE generalised estimation equations, WGEE weighted generalised estimating equations, MIGEE generalised estimating equations with multiple imputation, GLMM generalised linear mixed model, MCAR missing completely at random, MAR missing at random, MNAR missing not at random.

Three models which incorporated random effects were also fitted. The first was a standard generalised linear mixed model (GLMM), followed by shared parameter and pattern mixture models which were fit to the data using maximum likelihood estimation. As described in Chapter 6, the shared parameter model jointly modelled the outcome and time to exit from the study due to death or being permanently lost to follow-up through the specification of a shared random effect. The pattern mixture model broke down the participants into distinct groups based on the time at which they left the study and the effect of each covariate was allowed to vary within each group.

Four different outcomes were considered; anxiety and depression (both binary outcomes), activity level (a three level ordinal outcome) and disability level (a four level ordinal outcome). However, as in the previous Chapter the results for anxiety, which exhibited similar patters to those for other outcomes, have been omitted from this chapter and can be found in Appendix E. Unadjusted and multivariable adjusted

models were run for all outcomes and the results are presented in this chapter.

7.3 Dataset

The dataset used in analyses included data from 3617 participants in the (SLSR) who suffered a stroke between the 1st January 1995 and the 31st December 2007. The baseline characteristics of the participants are summarised in Table 7.2. All characteristics displayed in Table 7.2 were previously described in Chapter 4 and selected for inclusion in the study as they were thought to be potential predictors of poor outcome and dropout. They were also complete for the majority of participants and provided a mix of binary, ordinal, nominal and continuous independent variables.

There were some missing data within these variables, most notably in disability level, measured using the Barthel Index recorded 7-10 days post stroke, which was unknown for 25% of all SLSR participants. This was missing in the majority of cases due to deaths which occurred shortly after the stroke event. Ethnicity was unknown in 2.5% of participants and level of consciousness, measured using the Glasgow Coma Score [GCS], in 3.9%. Stroke subtype, as previously described in Chapter 4, includes an unknown/unclassified category. The unknown category represents a distinct group and includes participants for which it was not possible to identify a specific subtype. Stroke had to have been confirmed for all participants in order to be eligible for inclusion in the SLSR, but for some it was not possible to determine the type of stroke, despite having scans and diagnostic tests. Therefore unknown subtype is not considered to be missing data.

Only participants who contributed follow-up data in the first five years after stroke were included in the models presented in this chapter. In total 2288 patients were eligible for inclusion. Among these participants Barthel Index was missing for 6.2%,

Table 7.2: Baseline characteristics of all SLSR participants (1995-2007), and of those with ≥1 complete follow-up

	All participants	Participants with ≥1 follow-up
Total, N	3617	2288
Age, mead(sd)	70.3(14.7)	68.4(14.0)
Missing, n(%)	*0*	*0*
Gender, n(%)		
Male	1833(50.7)	1225(53.5)
Female	1784(49.3)	1063(46.5)
Missing, n(%)	*0*	*0*
Ethnicity, n(%)		
White	2622(74.4)	1596(70.6)
Black	702(19.9)	517(22.9)
Other	201(5.7)	148(6.6)
Missing, n(%)	*92(2.5)*	*28(1.2)*
Glasgow coma score, median (IQR)	15(11-15)	15(14-15)
Missing, n(%)	*142(3.9)*	*66(2.9)*
Barthel Index at 7-10 days, n(%)		
Moderate-severe disability	1590(58.6)	1079(50.3)
Mild disability	465(17.2)	438(20.4)
Independent	657(24.2)	629(29.3)
Missing, n(%)	*905(25.0)*	*142(6.2)*
Stroke Subtype, n(%)		
Ischaemic	2614(72.3)	1818(79.5)
PICH	486(13.4)	242(10.6)
SAH	199(5.5)	99(4.3)
Unknown	318(8.8)	129(5.6)
Missing	*0*	*0*
Missing at least one characteristic, n(%)	*1005(27.8)*	*201(8.9)*

Abbreviations: sd standard deviation, IQR inter quartile range, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

GCS for 2.9% and 1.2% had unknown ethnicity. Across all baseline variables, 27.8% of all participants had one or more missing values, while the rate was 8.9% in those eligible for inclusion in the models. Participants with missing data were excluded from the models. While missing covariate information may also have an impact, this was not explored here and analyses focus on the impact of missing outcome data.

Some participants did not complete certain items during a follow-up despite taking part in a face to face interview or partially completing and returning a postal questionnaire. This resulted in varying numbers of participants being included in the models for each of the four outcomes. Missing data in individual outcomes are further described in the following sections prior to the results from the models.

7.4 Comparison of models exploring the association between baseline characteristics and post stroke depression

Baseline characteristics associated with depression were also explored using the same models as for anxiety. Depression was measured using the depression subscale of the Hospital Anxiety and Depression Scale and is often categorised in analysis. The binary form which identified participants with possible or probable depression was used in this section.

The completeness of the depression data, along with rates of depression, at each time point are summarised in Table 7.3. Depression measurements were only available for between 36.2% and 59.3% of participants alive at each time point, with 1877 providing at least one HADS-D measurement. Up to 40% of the missing data at the three month and two year follow-ups were due to the scale not being included on the data collection form. Up to one quarter of participants who did complete the follow-up did not complete the HADS-D, despite it being included on the data collection form.

Table 7.3: Completeness of HADS Depression measurements and prevalence of depression in SLSR participants (1995-2007)

	3 months	1 year	2 years	3 years	4 years	5 years
Total alive, N	2615	2320	2123	1917	1782	1654
Completed depression measurement, n(%)	947(36.2)	1076(46.4)	842(39.7)	1137(59.3)	978(55.4)	814(50.3)
Not depressed	640(67.6)	767(71.3)	585(69.5)	779(68.5)	672(68.7)	564(69.3)
Depressed	307(32.4)	309(28.7)	257(30.5)	358(31.5)	306(31.3)	250(30.7)
Reason for missing measurement, n(%)						
Lost to follow-up	783(46.9)	595(47.8)	534(41.7)	571(73.2)	578(73.3)	627(77.8)
HADS not on form	609(36.5)	371(29.8)	521(40.7)	0	0	0
HADS not done for other reason	278(16.8)	278(22.4)	226(17.6)	209(26.8)	211(26.7)	179(22.2)

Table shows parameter estimates from logistic GLMMs with random intercept for the relationship between depression and time since stroke. Models were adjusted for age, sex, ethnicity, stroke subtype, Glasgow coma score and disability 7-10 days after stroke.

7.4.1 Handling of time in models for depression

Depression was a binary outcome and so analysed using logistic models. Prior to conducting the main analyses, the relationship between time and depression was explored in a logistic GLMM (Table 7.4). There did not appear to be any significant change over time in the likelihood of being depressed, and no evidence of a nonlinear relationship. Time was therefore included as a linear covariate in the models.

Table 7.4: Relationship between time after stroke and depression

	beta	se	t	p-value
time	-0.001	0.025	-0.04	0.97
time	-0.04	0.093	-0.43	0.666
$time^2$	0.008	0.018	0.44	0.662
time	-0.372	0.241	-1.54	0.123
$time^2$	0.169	0.11	1.54	0.123
$time^3$	-0.021	0.014	-1.49	0.136

Table shows parameter estimates from logistic GLMMs with random intercept for the relationship between depression and time since stroke. Models were adjusted for age, sex, ethnicity, stroke subtype, Glasgow coma score and disability 7-10 days after stroke.

7.4.2 Unadjusted logistic models for depression

The results of the unadjusted GEE approaches for depression are summarised in Table 7.5. The models were fairly consistent with the strength and direction of the relationships similar in the majority of the models. The GEE and WGEE models found an increased risk in those of other ethnicity when compared to white but the difference was not significant in the MIGEE model; although there was a slightly increased risk, the beta coefficient was lower than in the other models. All models, except for the mortal WGEE model, also found an increase in the odds of depression with increasing level of disability. The standard errors estimated from the MIGEE models were greater than those from the other models. However, as described in Chapter 5 Section 5.3.3.2, as WGEEs are only applicable in situations with monotone missing data patters, intermittent missing values were imputed to allow for the weighting of data to account for missingness due to drop out and therefore the standard errors are likely to be underestimated.

Table 7.5: Unadjusted logistic GEE approach models exploring the associations between baseline characteristics and post stroke depression

	GEE				WGEE (mortal)				WGEE (immortal)				MIGEE			
	beta	se	z	p-value	beta	se	z	p-value	beta	se	z	p-value	beta	se	t	p-value
Age	0.000	0.003	-0.09	0.93	-0.002	0.002	-1.00	0.318	-0.001	0.002	-0.67	0.502	-0.005	0.006	-0.79	0.439
Sex																
Male	ref				ref				ref				ref			
Female	0.070	0.086	0.80	0.421	0.038	0.055	1.00	0.491	0.041	0.052	0.73	0.468	0.047	0.143	0.33	0.745
Ethnicity																
White	ref				ref				ref				ref			
Black	0.122	0.099	1.23	0.218	0.113	0.07	1.61	0.107	0.099	0.072	1.37	0.171	0.203	0.162	1.25	0.222
Other	0.363	0.154	2.36	0.018	0.440	0.119	3.72	<0.001	0.398	0.123	3.24	0.001	0.113	0.279	0.40	0.690
Subtype																
Infarct	ref				ref				ref				ref			
PICH	0.133	0.134	1.00	0.319	0.15	0.087	1.72	0.085	0.164	0.09	1.82	0.069	0.155	0.144	1.08	0.284
SAH	-0.140	0.206	-0.68	0.495	-0.180	0.12	-1.5	0.134	-0.202	0.128	-1.58	0.114	0.142	0.396	0.36	0.721
Undefined	-0.254	0.213	-1.19	0.234	-0.281	0.162	-1.73	0.084	-0.266	0.165	-1.61	0.107	-0.017	0.251	-0.07	0.946
GCS	-0.011	0.017	-0.62	0.536	-0.008	0.012	-0.67	0.503	-0.005	0.012	-0.42	0.674	0.013	0.023	0.57	0.574
7-10d Disability																
Severe	ref				ref				ref				ref			
Moderate	-0.016	0.131	-0.12	0.903	-0.025	0.089	-0.28	0.779	-0.029	0.09	-0.32	0.749	0.072	0.122	0.59	0.554
Mild	-0.107	0.118	-0.91	0.364	-0.154	0.084	-1.83	0.067	-0.160	0.085	-1.88	0.060	-0.114	0.138	-0.82	0.415
Indep	-0.516	0.110	-4.69	<0.001	-0.636	0.068	-9.35	<0.001	-0.695	0.069	-10.07	<0.001	-0.572	0.196	-2.92	0.008

All models were adjusted for time since stroke as a linear covariate.

Abbreviations: GEE generalised estimation equations, WGEE weighted generalised estimating equations, MIGEE generalised estimating equations with multiple imputation, se standard error, GCS Glasgow coma score, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

The findings of the unadjusted likelihood based models are presented in Table 7.6 with the full results from the pattern mixture models presented in Table 7.7. Prior to fitting these models the random effects structure was explored. Initially a random intercept model was fitted to the data which included all baseline characteristics of interest and time since stroke. A random slope was then added to the model and compared to the random intercept model using a likelihood ratio test. The random intercept+slope model was not a significantly better fit than the intercept only model ($\chi^2_1 = 1.78$, p=0.182). Therefore, all random effects models for depression included only a random intercept.

In the pattern mixture models there was again no evidence that the associations between baseline characteristics and depression differed by time of exit from the study.

Using the weighted averages of the coefficients from the pattern mixture models, the findings from the GLMMs, shared parameter and pattern mixture models were broadly similar (Table 7.6). As with the GEE approach models, all identified differences by level of disability. There were again discrepancies between the models regarding ethnicity and statistically significance; the beta coefficient for the difference between white and other ethnic groups in the shared parameter and pattern mixture models was larger than in the GLMM, with a statistically significant difference observed in the shared parameter model only.

The level of uncertainty surrounding the parameter estimates also differed across the three models; the standard errors were largest in the pattern mixture models, followed by the shared parameter models.

The variance of the shared random effects in each of the shared parameter models are also displayed in (Table 7.6). The variance was significantly different from

zero in all models. This suggests that there exists some underlying trait, not accounted for by the factors in the models, which drives increases in both the likelihood of exit from the study and the odds of being depressed.

Table 7.6: Unadjusted likelihood based logistic models exploring the association between baseline characteristics and post stroke depression

	GLMM				Shared parameter model							Pattern mixture			
									Shared random effect						
	beta	se	t	p-value	beta	se	t	p-value	Var	se	p-value	beta	se	t	p-value
Age	0.000	0.004	0.00	1.000	-0.004	0.005	-0.80	0.424	1.341	0.102	<0.001	-0.004	0.007	-0.57	0.569
Sex									1.340	0.097	<0.001				
Male	ref				ref							ref			
Female	0.095	0.112	0.85	0.395	0.012	0.14	0.09	0.928				-0.049	0.169	-0.29	0.772
Ethnicity									1.336	0.098	<0.001				
White	ref				ref							ref			
Black	0.176	0.131	1.34	0.180	0.181	0.169	1.07	0.285				0.186	0.192	0.97	0.332
Other	0.304	0.204	1.49	0.136	0.61	0.288	2.12	0.034				0.515	0.331	1.56	0.119
Subtype									1.340	0.096	<0.001				
Infarct	ref				ref							ref			
PICH	0.184	0.179	1.03	0.303	0.013	0.164	0.08	0.936				0.034	0.192	0.18	0.857
SAH	-0.191	0.263	-0.73	0.465	0.063	0.287	0.22	0.826				-0.132	0.295	-0.45	0.653
Undefined	-0.335	0.254	-1.32	0.187	-0.086	0.587	-0.15	0.881				-0.095	0.521	-0.18	0.857
GCS	-0.011	0.022	-0.50	0.617	-0.014	0.027	-0.52	0.603	1.339	0.096	<0.001	-0.011	0.0300	-0.37	0.711
7-10d Disability									1.297	0.095	<0.001				
Severe	ref				ref							ref			
Moderate	-0.209	0.166	-1.26	0.208	0.009	0.219	0.04	0.968				-0.012	0.222	-0.05	0.96
Mild	-0.130	0.153	-0.85	0.395	-0.159	0.199	-0.80	0.424				0.006	0.256	0.02	0.984
Indep	-0.655	0.139	-4.71	<0.001	-0.732	0.174	-4.21	<0.001				-0.618	0.201	-3.07	0.002

All models were adjusted for time since stroke as a linear covariate.

Abbreviations: GLMM generalised linear mixed model, se standard error, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

Table 7.7: Unadjusted logistic pattern mixture models for the association between baseline characteristics and post stroke depression (part 1 of 2)

	Overall				max time = 3m				max time = 1yr			
	beta	se	t	p-value	beta	se	t	p-value	beta	se	t	p-value
Age	-0.006	0.006	-1.00	0.317	-0.032	0.027	-1.19	0.234	0.012	0.018	0.67	0.503
Sex												
Male	ref				ref				ref			
Female	0.010	0.159	0.06	0.952	-1.639	0.808	-2.03	0.042	0.102	0.493	0.21	0.834
Ethnicity												
White	ref				ref				ref			
Black	0.135	0.189	0.71	0.478	0.099	0.813	0.12	0.904	-0.025	0.524	-0.05	0.96
Other	0.521	0.303	1.72	0.086	-0.125	0.933	-0.13	0.897	0.006	0.869	0.01	0.992
Subtype												
Infarct	ref				ref				ref			
PICH	0.033	0.171	0.19	0.849	0.013	0.652	0.02	0.984	-0.068	0.524	-0.13	0.897
SAH	-0.100	0.287	-0.35	0.726	-0.003	0.936	0.00	1.000	0.036	0.789	0.05	0.96
Undefined	-0.078	0.521	-0.15	0.881	0.123	1.532	0.08	0.936	-0.327	1.237	-0.26	0.795
GCS	-0.034	0.03	-1.13	0.259	-0.021	0.141	-0.15	0.881	0.031	0.114	0.27	0.787
7-10d Disability												
Severe	ref				ref				ref			
Moderate	-0.012	0.216	-0.06	0.952	-0.024	0.965	-0.02	0.984	-0.025	0.875	0.03	0.976
Mild	-0.014	0.235	-0.06	0.952	0.022	0.97	0.02	0.984	0.037	0.935	0.16	0.873
Indep	-0.633	0.187	-3.39	0.001	-0.001	0.725	0.00	1.000	0.103	0.690	0.13	0.897

Pattern mixture model showing the association between covariates and depression in those with complete data (the main effect) and the additional effect the parameters have in those who dropped out broken down by the maximum follow-up time of those without five years of follow-ups.

Abbreviations: se standard error, GCS Glasgow coma score, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

Table 7.14 Unadjusted logistic pattern mixture models for the association between baseline characteristics and post stroke depression (part 2 of 2)

	max time = 2yr				max time = 3yr				max time = 4yr			
	beta	se	t	p-value	beta	se	t	p-value	beta	se	t	p-value
Age	0.041	0.016	2.56	0.011	0.011	0.012	0.92	0.358	-0.008	0.011	-0.73	0.465
Sex												
Male	ref				ref				ref			
Female	0.295	0.402	0.73	0.465	0.307	0.34	0.90	0.368	0.244	0.295	0.83	0.407
Ethnicity												
White	ref				ref				ref			
Black	0.124	0.458	0.27	0.787	0.096	0.393	0.24	0.81	0.103	0.315	0.33	0.741
Other	0.045	0.757	0.06	0.952	0.025	0.699	0.04	0.968	-0.005	0.588	-0.01	0.992
Subtype												
Infarct	ref				ref				ref			
PICH	0.013	0.413	0.03	0.976	0.054	0.325	0.17	0.865	0.000	0.326	0.00	1.000
SAH	-0.002	0.660	0.00	1.000	-0.265	0.548	-0.48	0.631	-0.004	0.522	-0.01	0.992
Undefined	0.013	0.924	0.01	0.992	0.036	0.895	0.04	0.968	0.025	0.896	0.03	0.976
GCS	0.08	0.088	0.91	0.363	0.045	0.057	0.79	0.430	0.045	0.057	0.79	0.430
7-10d Disability												
Severe	ref				ref				ref			
Moderate	0.024	0.757	0.03	0.976	-0.003	0.633	0.000	1000	0.025	0.510	0.05	0.960
Mild	0.104	0.757	0.14	0.889	-0.079	0.593	-0.13	0.897	0.076	0.532	0.14	0.889
Indep	0.009	0.521	0.02	0.984	0.012	0.400	0.03	0.976	-0.003	0.357	-0.01	0.992

Pattern mixture model showing the association between covariates and depression in those with complete data (the main effect) and the additional effect the parameters have in those who dropped out broken down by the maximum follow-up time of those without five years of follow-ups.

Abbreviations: se standard error, GCS Glasgow coma score, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

The estimates from the likelihood based models were transformed into marginal estimates and compared to the results from the GEE approach models in Figure 7.1. There was good agreement between the GEE approach models and random effect based models. The conclusions drawn from the models regarding the direction, strength and significance of the associations between the covariates in the models and odds of depression were the same in the majority of cases.

As a result of the larger standard errors in some models, the confidence intervals for the odds ratios estimated using MIGEE, shared parameter and pattern mixture models were wider than for any of the other models.

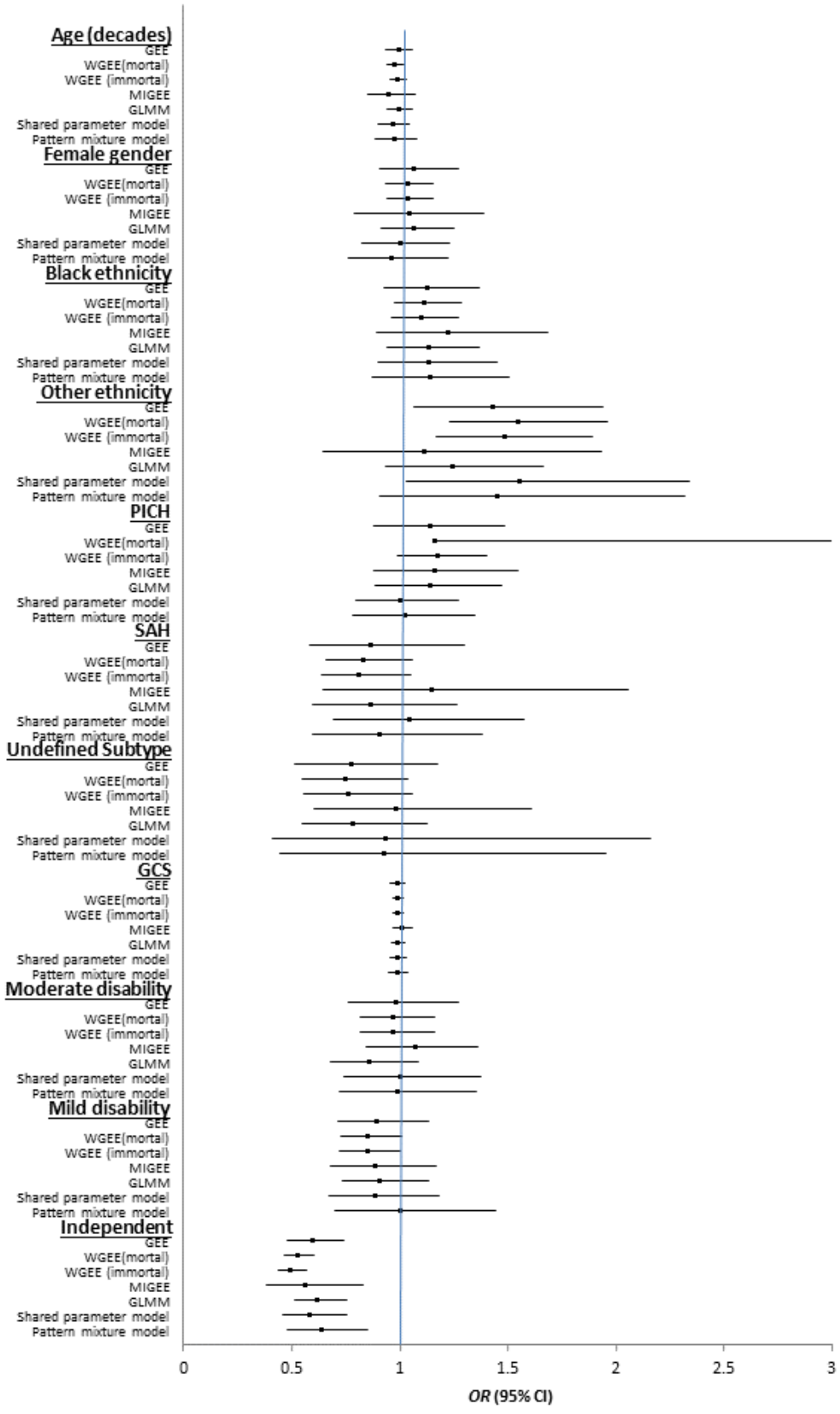


Figure 7.1: Unadjusted estimates of the marginal effect of baseline characteristics on the odds of post stroke depression

7.4.3 Adjusted logistic models for depression

The depression analysis was also repeated using models which adjusted for other baseline characteristics. The results of the GEE approach models are in Table 7.8 and the likelihood based models in Table 7.9. The findings from both sets of models were very similar to those observed for the unadjusted models described in the previous section. Significant differences were observed only by disability level at 7-10 days after stroke and, in some models, between other and white ethnic groups. Standard errors from the MIGEE model were larger than for the other GEE approach models and the shared parameter model produced larger standard errors than the random effects model. The variance of the shared random effect remained significant in the adjusted model, again suggesting the presence of an underlying MNAR process.

The results of both the GEE approach models and the marginal estimates from the random effects based models are also summarised in Figure 7.2. This figure further highlights the consistency between the models with only a few slight differences observed. All models apart from the MIGEE suggested a slightly reduced risk of depression in those with an undefined stroke subtype relative to infarcts, though this was not significant; the point estimate in the MIGEE model suggested a possibly increased risk, though again this was not significant. Again, the conclusions drawn were relatively consistent across all models, with large overlap in confidence intervals even where differences were observed.

Table 7.8: Adjusted logistic GEE approach models exploring the associations between baseline characteristics and post stroke depression

	GEE				WGEE (mortal)				WGEE (immortal)				MIGEE			
	beta	se	z	p-value	beta	se	z	p-value	beta	se	z	p-value	beta	se	t	p-value
Age	-0.002	0.004	-0.59	0.566	-0.006	0.002	-2.71	0.007	-0.004	0.002	-2.1	0.036	-0.006	0.006	-1.10	0.278
Sex																
Male	ref				ref				ref				ref			
Female	0.049	0.088	0.57	0.579	0.03	0.054	0.55	0.58	0.023	0.055	0.41	0.681	0.032	0.133	0.24	0.812
Ethnicity																
White	ref				ref				ref				ref			
Black	0.105	0.105	0.99	0.316	0.048	0.072	0.67	0.505	0.042	0.074	0.57	0.567	0.127	0.155	0.82	0.42
Other	0.352	0.16	2.26	0.028	0.342	0.128	2.68	0.007	0.288	0.133	2.17	0.030	0.157	0.281	0.56	0.581
Subtype																
Infarct	ref				ref				ref				ref			
PICH	0.028	0.14	0.15	0.842	0.057	0.091	0.62	0.532	0.036	0.094	0.38	0.701	0.083	0.141	0.59	0.557
SAH	-0.232	0.215	-1.01	0.280	-0.164	0.132	-1.24	0.215	-0.175	0.14	-1.25	0.211	-0.063	0.402	-0.17	0.868
Undefined	-0.199	0.215	-0.99	0.354	-0.147	0.155	-0.95	0.342	-0.121	0.157	-0.77	0.441	-0.114	0.255	-0.45	0.656
GCS	0.017	0.02	0.92	0.382	0.012	0.013	0.87	0.382	0.02	0.014	1.43	0.153	0.04	0.026	1.53	0.136
7-10d Disability																
Severe	ref				ref				ref				ref			
Moderate	-0.042	0.135	-0.32	0.756	-0.082	0.096	-0.85	0.395	-0.023	0.097	-0.24	0.81	0.045	0.123	0.37	0.714
Mild	-0.132	0.124	-1.06	0.289	-0.149	0.091	-1.64	0.101	-0.159	0.092	-1.73	0.084	-0.161	0.141	-1.14	0.260
Indep	-0.545	0.119	-4.60	<0.001	-0.711	0.072	-9.85	<0.001	-0.729	0.073	-10.05	¡0.001	-0.625	0.214	-2.91	0.009

All models were adjusted for time since stroke as a linear covariate.

Abbreviations: GEE generalised estimation equations, WGEE weighted generalised estimating equations, MIGEE generalised estimating equations with multiple imputation, se standard error, GCS Glasgow coma score, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

Table 7.9: Adjusted likelihood based logistic models exploring the association between baseline characteristics and post stroke depression

	GLMM				Shared parameter model						
									Shared random effect		
	beta	se	t	p-value	beta	se	t	p-value	Var	se	p-value
Age	-0.003	0.004	-0.65	0.516	-0.006	0.006	-0.82	0.414	1.291	0.095	<0.001
Sex											
Male	ref				ref						
Female	0.077	0.112	0.69	0.49	0.148	0.144	1.18	0.237			
Ethnicity											
White	ref				ref						
Black	0.149	0.133	1.12	0.263	0.181	0.170	1.07	0.287			
Other	0.338	0.203	1.65	0.100	0.611	0.288	2.12	0.034			
Subtype											
Infarct	ref				ref						
PICH	0.043	0.181	0.24	0.809	0.068	0.233	0.29	0.772			
SAH	-0.289	0.272	-1.06	0.289	-0.272	0.336	-0.81	0.418			
Undefined	-0.310	0.250	-1.24	0.216	-0.431	0.301	-1.43	0.153			
GCS	0.025	0.024	1.04	0.297	0.030	0.031	0.97	0.332			
7-10d Disability											
Severe	ref				ref						
Moderate	-0.065	0.175	-0.37	0.712	-0.050	0.227	-0.22	0.826			
Mild	-0.165	0.163	-1.01	0.314	-0.176	0.211	-0.83	0.407			
Indep	-0.706	0.153	-4.61	<0.001	-0.773	0.191	-4.05	<0.001			

All models were adjusted for time since stroke as a linear covariate.

Abbreviations: GLMM generalised linear mixed model, se standard error, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

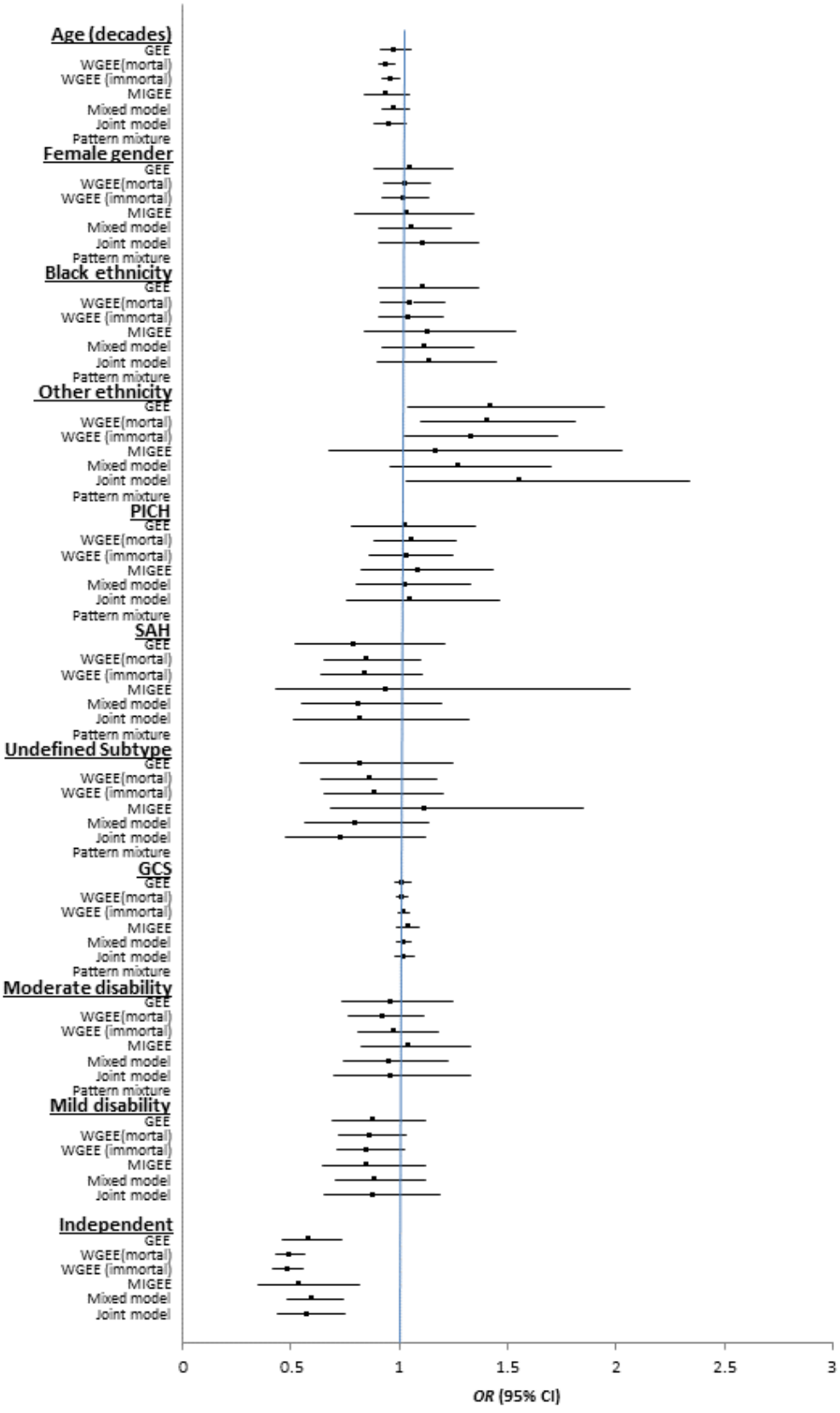


Figure 7.2: Adjusted estimates of the marginal effect of baseline characteristics on the odds of post stroke depression

7.5 Comparison of models exploring the association between baseline characteristics and post stroke activity level

The Frenchay Activities Index is used by the SLSR to measure activity levels. The Frenchay is often categorised into three groups; not active, moderately active and active. The distribution of participants in each of these groups, along with the completeness of the data, are summarised in Table 7.10.

Between 53.5% and 69.0% of those who were alive at each follow-up completed the Frenchay assessment. Among those for which no data were available, the majority were lost to follow-up and did not complete the interview at all. At each follow-up, between 5.8% and 13.0% of the missing data were due to incomplete Frenchay assessments in participants who otherwise completed the interview, at least in part. Overall 2233 participants contributed at least one Frenchay measurement.

Missing data were treated the same in the models regardless of the reason for being missing.

Table 7.10: Completeness of Frenchay Activities Index and prevalence of inactivity in SLSR participants (1995-2007)

	3 months	1 year	2 years	3 years	4 years	5 years
Total alive, N	2615	2320	2123	1917	1782	1654
Completed Frenchay assessment, n(%)	1638(62.6)	1600(69.0)	1136(53.5)	1261(65.8)	1105(62.5)	911(56.2)
Active	203(12.4)	293(18.3)	216(19.0)	212(16.8)	192(17.4)	170(18.7)
Moderately active	460(28.1)	492(30.8)	364(32.0)	422(33.5)	386(34.9)	303(33.3)
Not active	975(59.5)	815(50.9)	556(48.9)	627(49.7)	527(47.7)	438(48.1)
Reason for missing measurement, n(%)						
Lost to follow-up	856(87.4)	643(89.3)	930(94.2)	571(87.0)	578(87.3)	627(88.4)
Follow-up done - Frenchay not done	124(12.7)	77(10.7)	57(5.8)	85(13.0)	84(12.7)	82(11.6)

7.5.1 Handling of time and the proportional odds assumption in models for inactivity

Activity level was summarised using a three level ordinal variable. The proportional odds model is appropriate for modelling ordinal outcomes provided that the assumption that the effect of a covariate on the odds of being in outcome category k+1 relative to k is the same as on the odds of being in category k+2 relative to k+1. Before fitting the models to the activity data, the robustness of this assumption was explored. Proportional odds models were fitted to the data at each follow-up point and then the Brant test was used to determine whether there was any evidence that the covariates had a greater impact on the likelihood of being in one category over another.

The results of the test at one year after stroke are shown in Table 7.11; they were also typical of those observed at each of the other time points. None of the p-values were <0.05 and so there was no evidence to suggest violation of the proportional odds assumption. Therefore, it is likely that the assumption would hold in the random effects longitudinal models and so proportional odds models were used in the analysis of activity level.

Table 7.11: Brant test of proportional odds assumption in models for activity level at one year after stroke

Variable	chi2	df	p-value
All	10.59	10	0.39
Age	0.47	1	0.493
Sex	1.22	1	0.269
Black ethnicity	1.52	1	0.217
Other ethnicity	0.48	1	0.488
GCS	0.01	1	0.92
PICH	0.78	1	0.377
SAH	0.28	1	0.599
Unknown subtype	1.84	1	0.175
Mild disability	1.92	1	0.165
Moderate disability	1.23	1	0.267

Abbreviations: GCS Glasgow coma score, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

The association between time and activity level was explored in a series of proportional odds GLMMs. The estimated associations are summarised in Table 7.12. Initially a model assuming only a linear term was fitted, and then a quadratic, then a cubic and so on, until adding additional powers no longer improved the model fit significantly. The models suggested a cubic relationship existed and so terms for time, time squared and time cubed were included in all models for activity level.

To explore the random effects structure a random intercept proportional odds model was fitted to the data, which included all baseline characteristics of interest and time since stroke. Next, a random slope for the linear time component was added and there was no evidence to suggest that this random intercept+slope model was a better fit than the intercept only model (χ^2_1 =3.02, p=0.082). Therefore all random effects based models included a random intercept only.

Table 7.12: Relationship between time after stroke and inactivity

	beta	se	t	p-value
time	-0.027	0.012	-2.34	0.019
time	0.202	0.043	4.7	<0.001
$time^2$	-0.046	0.008	-5.54	<0.001
time	0.651	0.108	6	<0.001
$time^2$	-0.269	0.05	-5.37	<0.001
$time^3$	0.029	0.006	4.51	<0.001
time	1.082	0.228	4.74	<0.001
$time^2$	-0.661	0.189	-3.49	<0.001
$time^3$	0.149	0.056	2.64	0.008
$time^4$	-0.01	0.019	-0.53	0.298

Table shows parameter estimates from proportional odds GLMMs with random intercept for the relationship between activity level and time since stroke. Models were adjusted for age, sex, ethnicity, stroke subtype, Glasgow coma score and disability 7-10 days after stroke.

7.5.2 Unadjusted proportional odds models for activity level

Four proportional odds unadjusted GEE approach models were fitted to the data and the results are presented in Table 7.13. There was good agreement between the models in terms of identifying the factors associated with activity level. All GEE approach models found an increasing likelihood of being more inactive with increasing age. Females were less likely to be inactive than males, but only significantly so in the GEE and mortal WGEE models. Differences by ethnicity, stroke subtype, Glasgow coma score and disability level at 7-10 days after stroke were also significantly associated with level of activity.

Unlike in the logistic based models for anxiety and depression, the standard errors from the MIGEE models were similar to those from the other models.

Table 7.13: Unadjusted proportional odds GEE approach models exploring the associations between baseline characteristics and post stroke activity level

	GEE				WGEE (mortal)				WGEE (immortal)				MIGEE			
	beta	se	z	p-value	beta	se	z	p-value	beta	se	z	p-value	beta	se	t	p-value
Age	0.045	0.003	15.27	<0.001	0.039	0.003	11.97	<0.001	0.039	0.003	11.96	<0.001	0.029	0.004	7.25	<0.001
Sex																
Male	ref				ref				ref				ref			
Female	-0.156	0.078	-2.00	0.045	-0.101	0.078	-1.29	0.197	-0.134	0.080	-1.68	0.093	-0.063	0.093	-0.68	0.497
Ethnicity																
White	ref				ref				ref				ref			
Black	0.062	0.091	0.68	0.498	0.177	0.096	1.83	0.067	0.199	0.098	2.03	0.043	0.095	0.134	0.71	0.478
Other	0.76	0.175	4.34	<0.001	0.825	0.203	4.06	<0.001	0.81	0.209	3.88	<0.001	0.914	0.177	5.16	<0.001
Subtype																
Infarct	ref				ref				ref				ref			
PICH	0.109	0.128	0.85	0.397	0.126	0.109	-1.04	0.298	0.126	0.121	1.04	0.298	0.138	0.125	1.10	0.271
SAH	-1.042	0.177	-5.88	<0.001	-1.199	0.213	-5.63	<0.001	-1.199	0.213	-5.63	<0.001	-0.927	0.31	-2.99	0.003
Undefined	-0.346	0.167	-2.08	0.038	-0.25	0.243	-1.03	0.303	-0.25	0.243	-1.03	0.303	-0.546	0.177	-3.08	0.002
GCS	-0.077	0.017	-4.57	<0.001	-0.122	0.019	-6.45	<0.001	-0.129	0.019	-6.84	<0.001	-0.069	0.017	-4.06	<0.001
7-10d Disability																
Severe	ref				ref				ref				ref			
Moderate	-0.715	0.124	-5.77	<0.001	-1.245	0.112	-11.12	<0.001	-1.28	0.113	-11.34	<0.001	-0.594	0.112	-5.30	<0.001
Mild	-0.943	0.112	-8.44	<0.001	-1.603	0.111	-14.45	<0.001	-1.67	0.109	-15.29	<0.001	-0.943	0.112	-8.42	<0.001
Indep	-1.903	0.105	-18.05	<0.001	-2.401	0.103	-23.29	<0.001	-2.486	0.038	-23.92	<0.001	-2.096	0.107	-19.59	<0.001

All models were adjusted for a cubic relationship between time since and activity level.

Abbreviations: GEE generalised estimation equations, WGEE weighted generalised estimating equations, MIGEE generalised estimating equations with multiple imputation, se standard error, GCS Glasgow coma score, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

GLMMs and shared parameter proportional odds models were also fitted to the activity data and the results are reported in Table 7.14. Again, the two models produced very similar estimates and agreed with the GEE approach models in terms of the characteristics identified as being significantly associated with activity level. The standard errors were only slightly larger in the shared parameter model than in the random effects model. In the shared parameter models, the shared random effect variance was significant across all models, suggesting the presence of an underlying trait, not measured by the variables in the models, which results in some participants being more likely to be inactive and exit the study.

As with the models for depression, the coefficients and standard errors from the GLMMs and shared parameter models were transformed into estimates of the marginal effects; odds ratios and 95% confidence intervals are compared to those obtained from the GEE approach models in Figure 7.3.

The difference between the other and white ethnic groups was less in the GLMM and shared parameter model but otherwise the agreement across all the models was good. Confidence intervals were of similar width and apart from gender, all models agreed on the factors which were significantly associated with activity level.

Table 7.14: Unadjusted likelihood based proportional odds models exploring the association between baseline characteristics and post stroke activity level

	GLMM				Shared parameter model						
									Shared random effect		
	beta	se	t	p-value	beta	se	t	p-value	Var	se	p-value
Age	0.106	0.006	17.67	<0.001	0.097	0.007	13.86	<0.001	3.051	0.096	<0.001
Sex									3.331	0.105	<0.001
Male	ref				ref						
Female	-0.454	0.165	-2.75	0.006	-0.345	0.175	-1.97	0.049			
Ethnicity									3.321	0.104	<0.001
White	ref				ref						
Black	-0.032	0.197	-0.16	0.873	0.121	0.213	0.57	0.569			
Other	0.957	0.317	3.02	0.003	0.965	0.322	3.00	0.003			
Subtype									3.283	0.103	<0.001
Infarct	ref				ref						
PICH	0.109	0.267	0.41	0.682	0.024	0.275	0.09	0.928			
SAH	-2.678	0.386	-6.94	<0.001	-2.541	0.401	-6.34	<0.001			
Undefined	-0.751	0.351	-2.14	0.032	-0.822	0.374	-2.20	0.028			
GCS	-0.155	0.032	-4.84	<0.001	-0.091	0.040	-2.28	0.023	3.313	0.104	<0.001
7-10d Disability									2.911	0.093	<0.001
Severe	ref				ref						
Moderate	-1.336	0.232	-5.76	<0.001	-1.567	0.255	-6.15	<0.001			
Mild	-1.970	0.213	-9.25	<0.001	-2.023	0.225	-8.99	<0.001			
Independent	-3.967	0.201	-19.74	<0.001	-3.874	0.219	-17.69	<0.001			

All models were adjusted for a cubic relationship between time since and activity level.

Abbreviations: GLMM generalised linear mixed model, se standard error, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

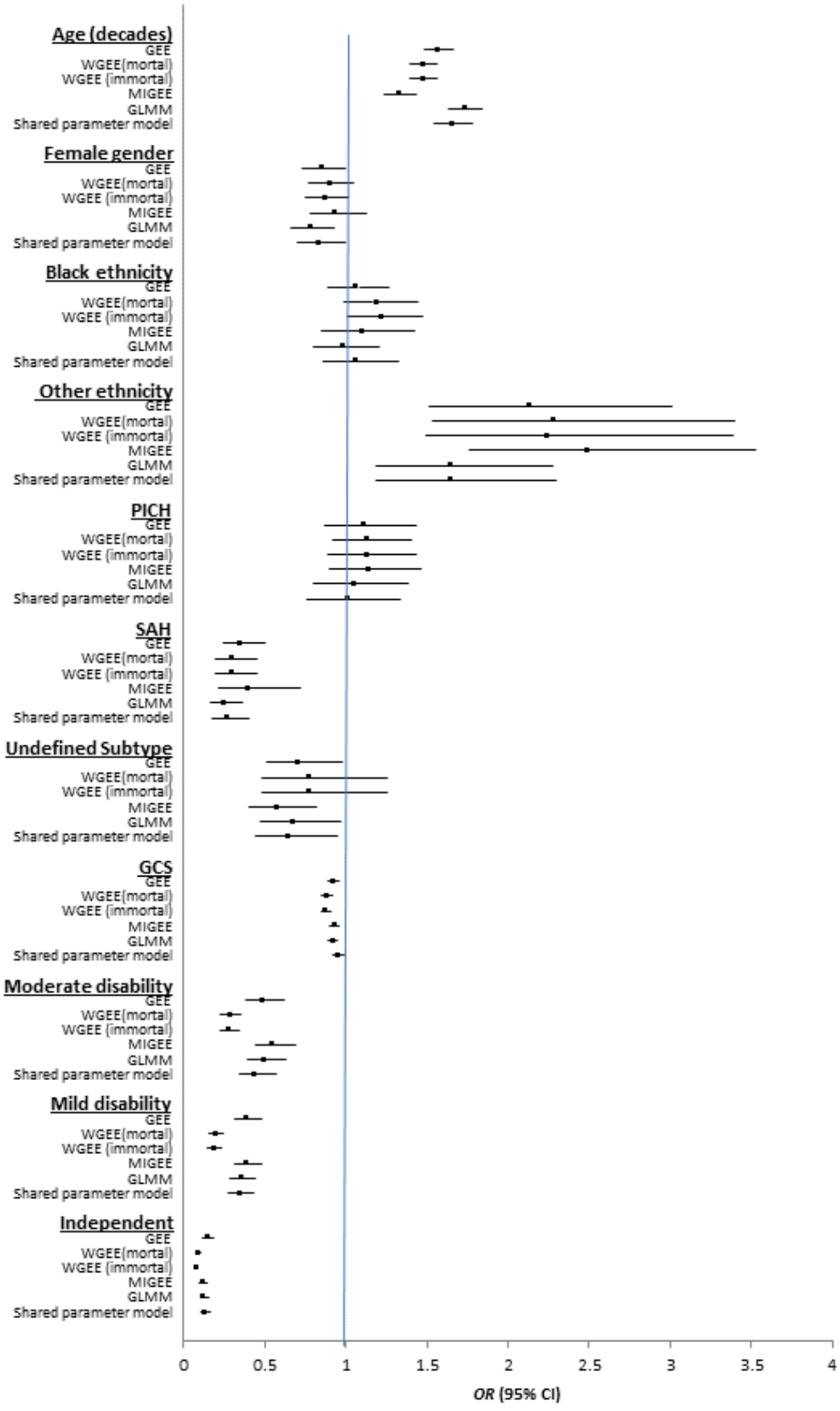


Figure 7.3: Unadjusted estimates of the marginal effect of baseline characteristics on the odds of increase level of inactivity

7.5.3 Adjusted proportional odds models for activity level

The analysis of activity level was repeated using multivariable models which adjusted estimates for the other baseline characteristics as well as time since stroke. The results of the GEE approaches are presented in Table 7.15 and the GLMM and shared parameter models are summarised in Table 7.16. The marginal odds ratios and 95% confidence intervals are also summarised in Figure 7.4.

All models led to the same interpretations regarding associations between baseline covariates and level of activity. The same factors, i.e. age, other ethnicity, GCS and Barthel Index at 7-10 days post stroke, identified as significant predictors in the unadjusted models were also significant in the adjusted models. In the unadjusted models no differences were observed between black and white ethnic groups, however after adjustment, increasing odds of inactivity were found in all models for black and other ethnic groups compared to white.

The point estimates of the odds ratios for all models were very similar overall and, as with the unadjusted models, there were only slight differences in the standard errors from different models.

Overall, in the analysis of activity level after stroke using proportional odds models, the choice of model, and the underlying assumption regarding the missing data mechanism, made no difference to the estimated associations between baseline characteristics and level of activity.

In the shared parameter model, the shared random effect variance remained significant in the adjusted model, again suggesting the presence of a MNAR mechanism.

Table 7.15: Adjusted proportional odds GEE approach models exploring the associations between baseline characteristics and post stroke activity level

	GEE				WGEE (mortal)				WGEE (immortal)				MIGEE			
	beta	se	z	p-value	beta	se	z	p-value	beta	se	z	p-value	beta	se	t	p-value
Age	0.047	0.003	13.69	<0.001	0.036	0.004	10.08	<0.001	0.037	0.004	10.53	<0.001	0.036	0.004	8.05	<0.001
Sex																
Male	ref				ref				ref				ref			
Female	-0.08	0.081	-0.99	0.323	-0.095	0.077	-1.23	0.220	-0.079	0.081	-0.99	0.324	0.03	0.095	0.32	0.752
Ethnicity																
White	ref				ref				ref				ref			
Black	0.417	0.097	4.30	<0.001	0.37	0.115	3.22	0.001	0.370	0.117	3.16	0.002	0.457	0.123	4.62	<0.001
Other	1.161	0.191	6.09	<0.001	1.374	0.194	7.08	<0.001	1.357	0.205	6.62	<0.001	1.272	0.271	4.70	<0.001
Subtype																
Infarct	ref				ref				ref				ref			
PICH	-0.113	0.136	-0.83	0.406	-0.267	0.158	-1.69	0.092	-0.296	0.151	-1.96	0.05	-0.263	0.129	-2.03	0.043
SAH	-0.513	0.208	-2.47	0.014	-1.158	0.283	-4.09	<0.001	-1.235	0.289	-4.27	<0.001	-1.067	0.367	-2.91	0.005
Undefined	-0.214	0.151	-1.42	0.156	-0.125	0.178	-0.70	0.481	-0.118	0.177	-0.07	0.503	-0.365	0.204	-1.79	0.075
GCS	0.009	0.019	0.53	0.598	0.015	0.019	0.78	0.434	0.04	0.019	2.12	0.034	0.025	0.019	1.31	0.192
7-10d Disability																
Severe	ref				ref				ref				ref			
Moderate	-0.89	0.127	-7.04	<0.001	-1.397	-0.106	-13.16	<0.001	-1.423	0.108	-13.15	<0.001	-0.790	0.122	-6.49	<0.001
Mild	-1.064	0.118	-8.98	<0.001	-1.731	0.108	-16.03	<0.001	-1.776	0.109	-16.34	<0.001	-1.10	0.123	-8.91	<0.001
Indep	-1.926	0.114	-16.92	<0.001	-2.377	0.130	-18.34	<0.001	-2.43	0.131	-18.53	<0.001	-2.215	0.120	-18.46	<0.001

All models were adjusted for time since stroke as a linear covariate.

Abbreviations: GEE generalised estimation equations, WGEE weighted generalised estimating equations, MIGEE generalised estimating equations with multiple imputation, se standard error, GCS Glasgow coma score, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

Table 7.16: Adjusted likelihood based proportional odds models exploring the association between baseline characteristics and post stroke activity level

	GLMM				Shared parameter model						
									Shared random effect		
	beta	se	t	p-value	beta	se	t	p-value	Var	se	p-value
Age	0.094	0.006	15.77	<0.001	0.101	0.012	8.33	<0.001	2.620	0.085	<0.001
Sex											
Male	ref				ref						
Female	-0.158	0.141	-1.12	0.263	-0.185	0.162	-1.14	0.259			
Ethnicity											
White	ref				ref						
Black	0.791	0.173	4.57	<0.001	0.81	0.192	4.22	<0.001			
Other	1.54	0.271	5.68	<0.001	1.652	0.306	5.4	<0.001			
Subtype											
Infarct	ref				ref						
PICH	-0.105	0.232	-0.45	0.653	-0.13	0.261	-0.5	0.617			
SAH	-0.877	0.34	-2.58	0.01	-0.95	0.354	-2.68	0.007			
Undefined	-0.426	0.301	-1.42	0.159	-0.441	0.309	-1.43	0.197			
GCS	0.026	0.03	0.86	0.387	0.048	0.031	1.55	0.254			
7-10d Disability											
Severe	ref										
Moderate	-1.637	0.229	-7.16	<0.001	-1.68	0.245	-6.86	<0.001			
Mild	-2.072	0.212	-9.77	<0.001	-2.549	0.219	-11.64	<0.001			
Independent	-3.621	0.204	-17.79	<0.001	-3.181	0.25	-12.72	<0.001			

All models were adjusted for time since stroke as a linear covariate.

Abbreviations: GLMM generalised linear mixed model, se standard error, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

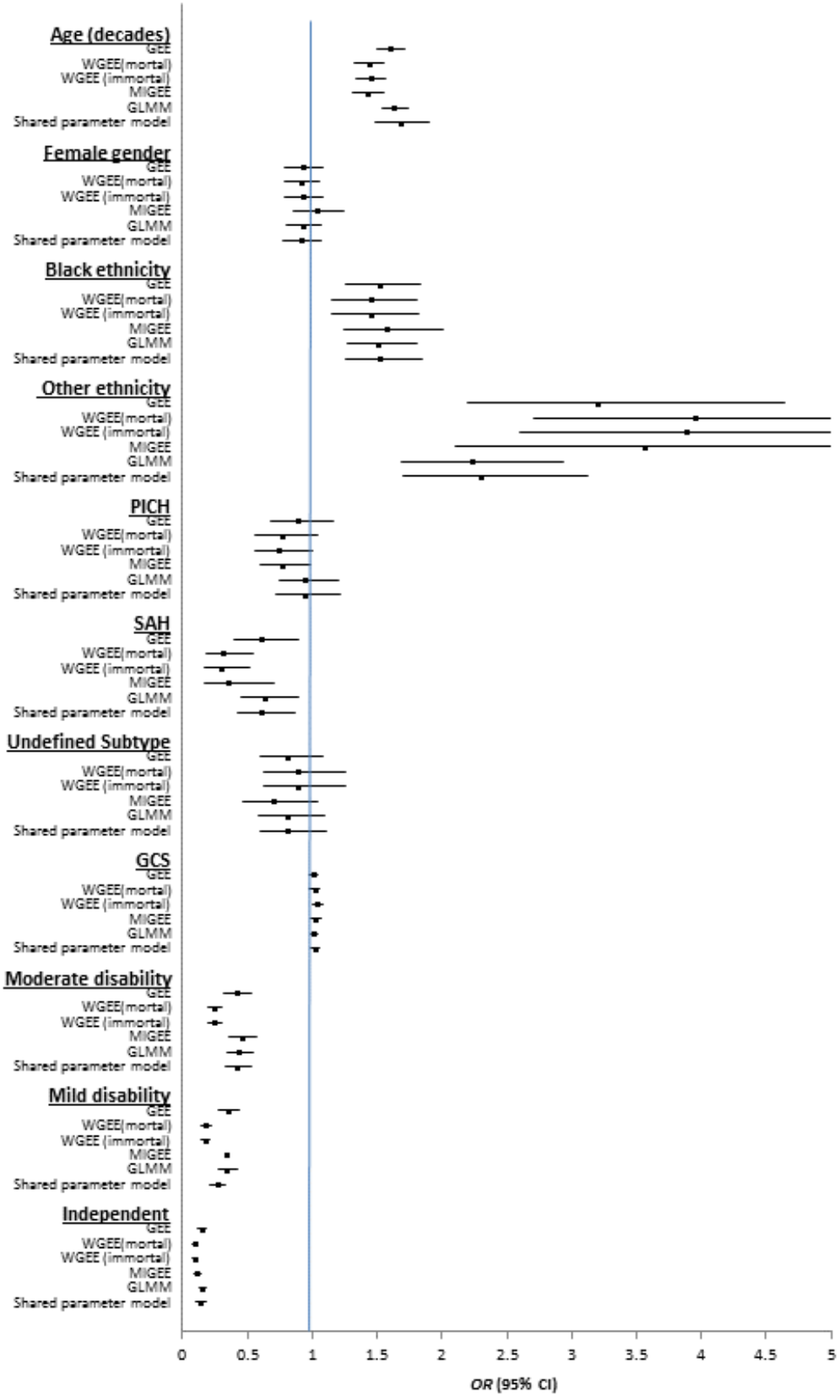


Figure 7.4: Adjusted estimates of the marginal effect of baseline characteristics on the odds of increase level of inactivity

7.6 Comparison of models exploring the association between baseline characteristics and post stroke disability

The final outcome considered was disability level, measured using the Barthel Index. The Barthel Index can by categorised to identify participants with severe, moderate and mild disability and those who are independent in activities of daily living. At follow-up, around one third of participants were independent and one third had mild disability. Among the remainder, approximately half were moderately disabled and the other half had severe disabilities (Table 7.17).

Between 55.4% and 71.2% of participants alive at each follow-up point completed the Barthel Index (Table 7.17). Missing Barthel scores in those who partially completed a follow-up were rarer than for other outcomes, accounting for <10% of the missing data. In total 2260 participants had at least one Barthel measurement.

Table 7.17: Completeness of Barthel Index and prevalence of disability in SLSR participants (1995-2007)

	3 months	1 year	2 years	3 years	4 years	5 years
Total alive, N	2615	2320	2123	1917	1782	1654
Completed BI measurement, n(%)	1751(66.9)	1659(71.2)	1176(55.4)	1294(67.5)	1146(64.9)	942(58.2)
Severe disability	331(18.9)	220(13.3)	141(12.0)	179(13.8)	141(12.3)	100(10.6)
Moderate disability	223(12.7)	201(12.1)	125(10.6)	139(10.7)	128(11.2)	111(11.8)
Mild disability	582(33.2)	578(34.8)	439(37.3)	471(36.4)	457(39.9)	364(38.6)
Independent	615(35.1)	660(39.8)	471(40.1)	505(39.0)	420(36.7)	367(39.0)
Reason for missing measurement, n(%)						
Lost to follow-up	856(98.7)	643(97.3)	930(98.2)	571(91.7)	578(93.1)	627(92.5)
Follow-up done - BI not done	11(1.3)	18(2.7)	17(1.8)	52(8.4)	43(6.9)	51(7.5)

Abbreviations: BI Barthel Index.

7.6.1 Handling of time and the proportional odds assumption in models for disability

Level of disability was summarised using a four point ordinal scale derived from the Barthel Index. Prior to fitting any models to the data the proportional odds assumption was explored. As with activity level, a series of Brant tests were applied after fitting proportional odds models at each follow-up time point. The results of the model at one year are shown in Table 7.18. The results at one year were typical of those observed at other time points with evidence that the proportional odds assumption did not hold. Therefore, the multinomial model was used in all analyses of level of disability.

Table 7.18: Brant test of proportional odds assumption in models for disability level at one year after stroke

Variable	chi2	df	p-value
All	65.36	20	<0.001
Age	21	2	<0.001
Sex	4.59	2	0.101
Black ethnicity	3.74	2	0.154
Other ethnicity	0.43	2	0.806
GCS	0.86	2	0.652
PICH	4.75	2	0.093
SAH	0.03	2	0.987
Unknown subtype	6.73	2	0.035
Mild disability	18.98	2	<0.001
Moderate disability	4.65	2	0.098

Abbreviations: GCS Glasgow coma score, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

The association between time and disability level was explored in a series of multinomial GLMMs with random slopes. The estimated associations are summarised in Table 7.19. There was evidence of a cubic relationship and so terms for time, time squared and time cubed were included in all models for disability level.

Table 7.19: Relationship between time and disability

	beta*	se	t	p-value
time	-0.074	0.011	-6.71	<0.001
time	0.1	0.045	2.25	0.025
$time^2$	-0.034	0.009	-4.02	<0.001
time	0.357	0.095	3.75	<0.001
$time^2$	-0.173	0.046	-3.74	<0.001
$time^3$	0.019	0.006	3.05	0.002
time	0.496	0.237	2.1	0.036
$time^2$	-0.302	0.207	-1.46	0.144
$time^3$	0.058	0.061	0.94	0.346
$time^4$	-0.004	0.006	-0.64	0.521

Table shows parameter estimates from multinomial GLMMs with random intercept for the relationship between disability level and time since stroke. Models were adjusted for age, sex, ethnicity, stroke subtype, Glasgow coma score and disability 7-10 days after stroke.

7.6.2 Unadjusted multinomial models for disability level

Two unadjusted models were applied to the Barthel data. The results from the multinomial GEE models are presented in Table 7.20 and the results of GLMMs in Table 7.21. A random intercept multinomial model was fitted to the data, which included all baseline characteristics of interest and time since stroke. The likelihood from the same model with a random slope for the linear time component was compared to the intercept only model and was not found to provide a better fit to the data (χ^2_1 =1.01, p=0.315). Therefore all random effects based models included a random intercept only. The odds ratios from the models, after transforming estimates from the random effects based models to marginal estimates, are also summarised in Figure 7.5.

Both models found differences by age and gender, with those of older age and female gender more likely to be more disabled. There were some differences between the estimated odds ratios from the two models. However, there were no obvious patterns with neither consistently producing higher or lower estimates than the other and the conclusions drawn regarding statistical significance were the same.

Table 7.20: Unadjusted multinomial GEE models exploring the associations between baseline characteristics and post stroke disability

	Mild				Moderate				Severe			
	beta	se	t	p-value	beta	se	t	p-value	beta	se	t	p-value
Age	0.028	0.003	9.77	<0.001	0.043	0.004	9.53	<0.001	0.088	0.005	16.06	<0.001
Sex												
Male	ref				ref				ref			
Female	0.528	0.079	6.68	<0.001	0.453	0.108	4.18	<0.001	0.860	0.111	7.72	<0.001
Ethnicity												
White	ref				ref				ref			
Black	-0.100	0.093	-1.09	0.278	-0.144	0.131	-1.100	0.269	-0.326	0.137	-2.38	0.017
Other	0.334	0.168	1.99	0.047	0.675	0.201	3.36	0.001	0.203	0.225	0.90	0.367
Subtype												
Infarct	ref				ref				ref			
PICH	-0.050	0.132	-0.38	0.704	0.197	0.166	1.19	0.234	0.056	0.175	0.32	0.747
SAH	-0.679	0.179	-3.80	<0.001	-1.686	0.323	-5.21	<0.001	-1.742	0.385	-4.53	<0.001
Undefined	0.034	0.162	0.21	0.832	-0.389	0.247	-1.57	0.116	-0.059	0.242	-0.24	0.808
GCS	-0.021	0.018	-1.18	0.237	-0.066	0.022	-2.98	0.003	-0.138	0.02	-6.83	<0.001
7-10d Disability												
Severe	ref				ref				ref			
Moderate	-0.141	0.128	-1.10	0.271	-0.484	0.158	-3.06	0.002	-1.688	0.18	-9.39	<0.001
Mild	-0.411	0.116	-3.55	<0.001	-1.201	0.153	-7.87	<0.001	-2.398	0.178	-13.46	<0.001
Independent	-1.250	0.107	-11.73	<0.001	-2.548	0.162	-15.69	<0.001	-3.267	0.182	-17.94	<0.001

All models were adjusted for a cubic relationship between time since and disability level.

Abbreviations: GEE generalised estimation equations, se standard error, GCS Glasgow coma score, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

Table 7.21: Unadjusted multinomial GLMMs exploring the association between baseline characteristics and post stroke disability

	Mild				Moderate				Severe			
	beta	se	t	p-value	beta	se	t	p-value	beta	se	t	p-value
Age	0.068	0.005	13.6	<0.001	0.079	0.006	13.17	<0.001	0.115	0.006	19.17	<0.001
Sex												
Male	ref				ref				ref			
Female	1.090	0.144	7.57	<0.001	1.024	0.154	6.65	<0.001	1.493	0.152	9.82	<0.001
Ethnicity												
White	ref				ref				ref			
Black	-0.300	0.171	-1.75	0.080	-0.322	0.184	-1.75	0.080	-0.394	0.181	-2.18	0.029
Other	0.331	0.274	1.21	0.226	0.628	0.286	2.20	0.028	0.381	0.285	1.34	0.180
Subtype												
Infarct	ref				ref				ref			
PICH	-0.031	0.232	-0.13	0.897	0.211	0.244	0.86	0.390	0.123	0.243	0.51	0.610
SAH	-1.734	0.345	-5.03	<0.001	-2.690	0.429	-6.27	<0.001	-2.308	0.391	-5.90	<0.001
Undefined	0.040	0.312	0.13	0.897	-0.398	0.351	-1.13	0.259	-0.139	0.334	-0.42	0.675
GCS	-0.072	0.027	-2.67	0.008	-0.157	0.031	-5.06	<0.001	-0.263	0.029	-9.07	<0.001
7-10d Disability												
Severe	ref				ref				ref			
Moderate	-0.648	0.206	-3.15	0.002	-1.009	0.216	-4.67	<0.001	-2.305	0.225	-10.24	<0.001
Mild	-1.215	0.189	-6.43	<0.001	-1.984	0.204	-9.73	<0.001	-3.266	0.215	-15.19	<0.001
Independent	-2.788	0.184	-15.15	<0.001	-4.080	0.209	-19.52	<0.001	-4.91	0.212	-23.16	<0.001

All models were adjusted for a cubic relationship between time since and disability level.

Abbreviations: GEE generalised estimation equations, se standard error, GCS Glasgow coma score, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

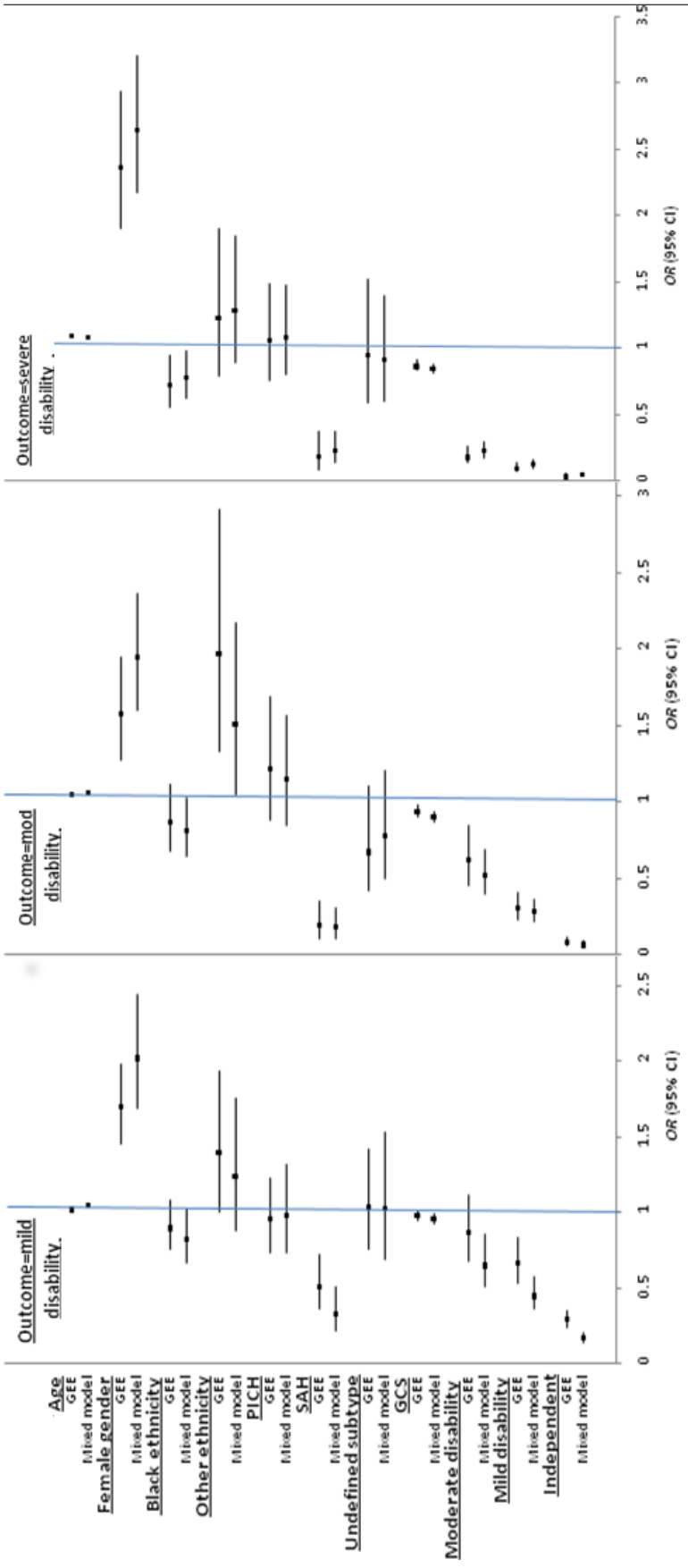


Figure 7.5: Unadjusted estimates of the marginal effect of baseline characteristics on the odds of disability

7.6.3 Adjusted multinomial models for disability level

Finally, multivariable multinomial models were applied to the disability data. The results of the multinomial GEE model are presented in Table 7.22, the multinomial GLMM in Table 7.23, and the corresponding marginal odds ratios displayed in Figure 7.6.

Overall the models agreed in terms of the effect of age, gender and disability level 7-10 days after stroke. However, there were some discrepancies in terms of ethnicity. In the GEE model, there were no statistically significant differences between black and white ethnic groups, though there was a trend towards increased risk in the black participants. In the GLMM, black participants were significantly less likely to have a disability than white participants. Further, in the GLMM differences between white and other ethnic groups were much smaller than those estimated by the GEE model, which found statistically significant differences. There was also a difference in the effect of Glasgow coma score at the time of stroke; in the GLMM participants with higher scores, i.e. those with greater levels of consciousness after stroke, were less likely to have moderate or severe disabilities, but no such differences were observed in the GEE model.

Table 7.22: Adjusted multinomial GEE model exploring the associations between baseline characteristics and post stroke disability

	Mild				Moderate				Severe			
	beta	se	t	p-value	beta	se	t	p-value	beta	se	t	p-value
Age	0.024	0.003	7.44	<0.001	0.039	0.005	7.60	<0.001	0.089	0.006	13.97	<0.001
Sex												
Male	ref				ref				ref			
Female	0.461	0.081	5.66	<0.001	0.349	0.118	2.96	0.003	0.500	0.127	3.94	<0.001
Ethnicity												
White	ref				ref				ref			
Black	0.044	0.098	0.45	0.653	0.125	0.144	0.87	0.384	0.330	0.162	2.04	0.041
Other	0.547	0.167	3.28	0.001	0.966	0.213	4.54	<0.001	0.833	0.242	3.45	0.001
Subtype												
Infarct	ref				ref				ref			
PICH	-0.131	0.135	-0.97	0.333	-0.005	0.182	-0.03	0.980	-0.186	0.205	-0.91	0.365
SAH	-0.436	0.204	-2.14	0.033	-1.217	0.365	-3.33	<0.001	-1.092	0.419	-2.61	0.009
Undefined	0.172	0.160	1.08	0.281	-0.141	0.251	-0.56	0.575	0.128	0.262	0.49	0.627
GCS	-0.042	0.013	-3.23	0.001	0.039	0.026	1.48	0.140	-0.016	0.025	-0.62	0.533
7-10d Disability												
Severe	ref				ref				ref			
Moderate	-0.365	0.135	-2.7	0.007	-0.791	0.168	-4.7	<0.001	-2.036	0.195	-10.43	<0.001
Mild	-0.647	0.124	-5.2	<0.001	-1.483	0.166	-8.94	<0.001	-2.570	0.189	-13.63	<0.001
Independent	-1.380	0.118	-11.66	<0.001	-2.652	0.180	-14.74	<0.001	-3.202	0.198	-16.17	<0.001

Model was adjusted for a cubic relationship between time since and disability level.

Abbreviations: GEE generalised estimation equations, se standard error, GCS Glasgow coma score, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

Table 7.23: Adjusted multinomial GLMM exploring the association between baseline characteristics and post stroke disability

	Mild				Moderate				Severe			
	beta	se	t	p-value	beta	se		p-value	beta	se		p-value
Age	0.05	0.005	10	<0.001	0.064	0.006	10.67	<0.001	0.109	0.006	18.17	<0.001
Sex												
Male	ref				ref				ref			
Female	0.748	0.125	5.98	<0.001	0.662	0.139	4.76	<0.001	0.942	0.141	6.68	<0.001
Ethnicity												
White	ref				ref				ref			
Black	-0.444	0.119	-3.73	<0.001	-0.351	0.162	-2.17	0.03	-0.401	0.174	-2.3	0.022
Other	0.301	0.265	1.14	0.254	0.452	0.251	1.8	0.072	0.391	0.247	1.58	0.114
Subtype												
Infarct	ref				ref				ref			
PICH	-0.379	0.203	-1.87	0.062	-0.413	0.221	-1.87	0.062	-0.688	0.225	-3.06	0.002
SAH	-1.077	0.304	-3.54	<0.001	-2.055	0.405	-5.07	<0.001	-1.815	0.385	-4.71	<0.001
Undefined	0.084	0.28	0.3	0.764	-0.278	0.332	-0.84	0.401	-0.348	0.336	-1.04	0.298
GCS	-0.04	0.024	-1.67	0.095	-0.087	0.029	-3	0.003	-0.184	0.027	-6.81	<0.001
7-10d Disability												
Severe	ref				ref				ref			
Moderate	-0.675	0.167	-4.04	<0.001	-1.204	0.201	-5.99	<0.001	-1.999	0.21	-9.52	<0.001
Mild	-1.312	0.154	-8.52	<0.001	-2.012	0.192	-10.48	<0.001	-3.176	0.201	-15.8	<0.001
Independent	-2.852	0.159	-17.94	<0.001	-3.995	0.197	-20.28	<0.001	-4.57	0.207	-22.08	<0.001

Model was adjusted for a cubic relationship between time since and disability level.

Abbreviations: GLMM generalised linear mixed model, se standard error, GCS Glasgow coma score, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

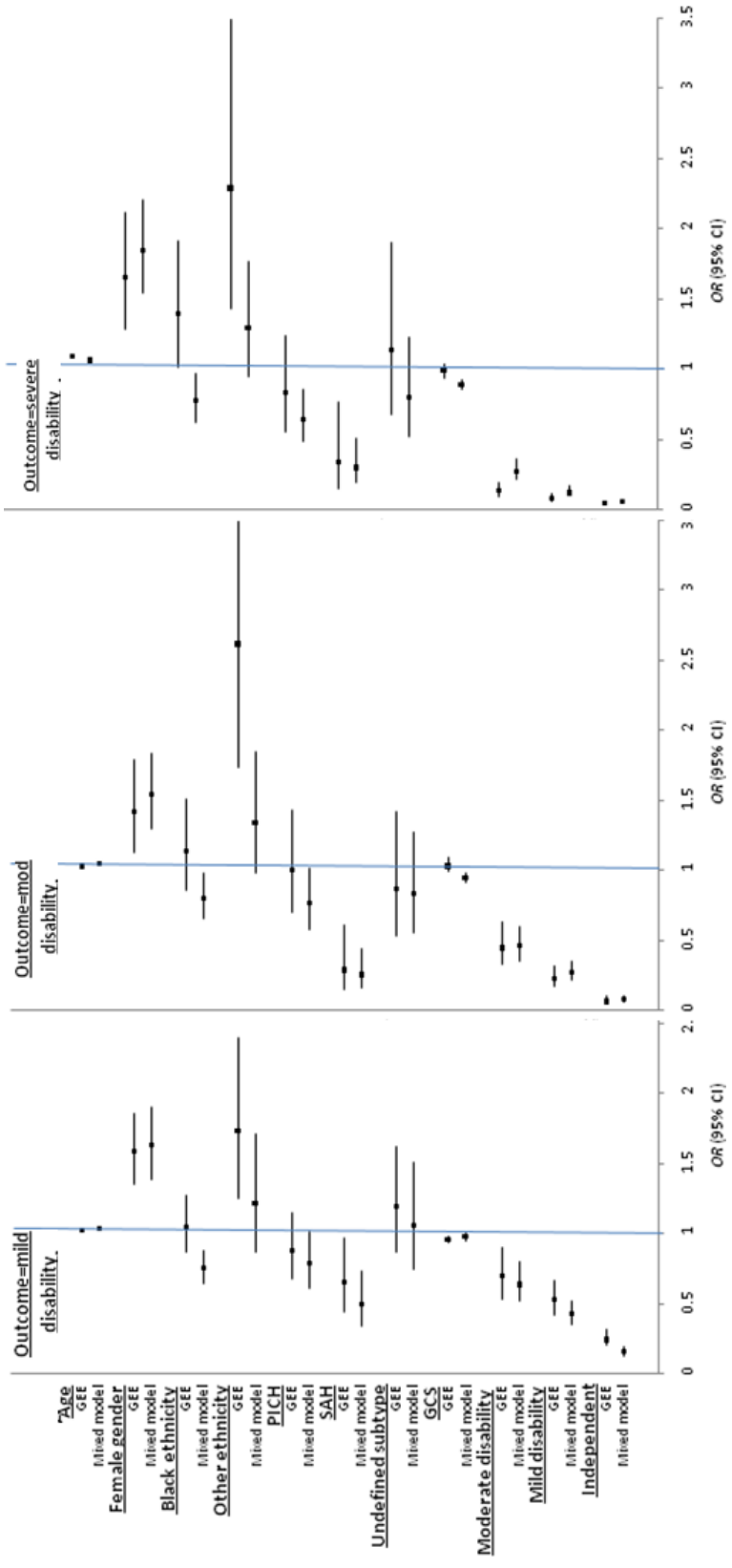


Figure 7.6: Adjusted estimates of the marginal effect of baseline characteristics on the odds of disability

7.7 Summary and conclusions

Models which assumed missing data were either MCAR, MAR or MNAR were used to explore the relationship between three outcomes and a number of baseline covariates.

Overall there was very good agreement between the models in terms of the factors identified as being significant predictors of outcome and in the point estimates of the associated odds ratios. Where differences were observed there were large overlaps in associated confidence intervals and none of the models consistently produced estimates that were higher or lower than all others.

In the logistic based models for depression the standard errors were greatest in the shared parameter and pattern mixture models. However, as described in Chapter 5 Section 5.3.4.3, a pooled standard error was used to combine the estimates across groups which does not take into account the fact that the proportions in each group, used in pooling, are themselves estimated, and so is likely to be underestimated. Similarly, in the WGEE analyses which require a monotone missing data pattern, intermittent missingness was imputed using data from the last and next completed follow-up. Although the results from Chapter 6 showed that a last observation carries forward imputation approach produced relatively unbiased estimates, imputing missing data using a single imputation approach results in standard errors which are too low. Multiple imputation of these data would prevent this, but this, but if employing multiple imputation for some outcomes data, it would be may be reasonable to use multiple imputation for all missing data as was done in the MIGEE models. Although there are limitations to these approaches, which in some cases produces unreliable standard error estimates, the estimates of the regression coefficients remain useful in detecting any biases arising from certain analytic approaches.

Shared parameter models were fitted for both depression and inactivity. In all

models (unadjusted and adjusted) the variance of shared random effect was significant. This suggests that there was some underlying trait not captured by the models which means that some participants are inherently more likely to exit the study and to have poorer outcomes. However, the parameters in the model were similar to the random effects model based on a MAR assumption. Therefore, although the drop-out from the study may be MNAR, the impact this had on associations between the baseline data included in the models and the outcomes of interest was minimal. This finding is in line with those from the pattern mixture models. In the pattern mixture model, the effect of the baseline characteristics were allowed to differ between groups which were defined according to time of exit from the study. There was no evidence that the effect of the coefficients differed across groups.

In conclusion, there was no evidence that the findings from the simpler, routinely applied models, such as a GEE approach or a GLMM, were affected by the potentially MNAR missing outcome data in the SLSR. However, the MNAR models applied in this study themselves rely on assumptions about the nature of the missing data mechanism. The shared parameter model makes the assumption that the outcome is independent of the missingness given the random effect. A simple shared random effect structure was assumed in this study with the drop-out process and the outcome model sharing a random intercept. There could potentially be further latent traits not accounted for by this model. For example, a shared random slope would allow for the possibility that there was an underlying process driving a deterioration in health as measured by the outcomes studied which also leads to the participant dropping out sooner. The structure of the models used in this study was somewhat limited by the fact that fitting these models is highly computationally intensive. The missingness process was modelled using a discrete time survival model which adjusted only for age and sex. Increasing the number of parameters and complexity of random effects structure lead to programs which would run for several days and fail to converge. Therefore, it is possible that the conditional independence of the

missing data and outcome processes does not hold.

Pattern mixture models also have limitations. To fit the pattern mixture model groups were defined by time of exit from the study and the effect of baseline variables allowed to differ by means of including an interaction between each variable and an indicator for group. Where there are many groups and several variables in the model, the number of parameters in the model can be high, again leading to potential computational difficulties. It was only possible to fit these models for the unadjusted models for depression. The pattern mixture approach assumes that the distribution of the outcome in the complete data is known conditional on the missing data pattern. In the analyses presented in this chapter groups were defined according to time of exit from the study, be that death or drop-out. It may be that the effect of the baseline variables on outcome differs between those who died and those who dropped out, but it was not possible to fit models which stratified participants into groups according to time of drop out and time of death. An additional limitation of both the MNAR used is that they allowed for potential MNAR drop out mechanisms but intermittent missingness was not considered in either.

The results from this study are discussed in detail in Chapter 8.

Chapter 8

Discussion

8.1 Summary of findings

The aim of this thesis was to compare and determine the most appropriate methods for handling non-continuous missing data in a cohort study with multiple follow-up assessments, namely the South London Stroke Register (SLSR). There was evidence found in exploratory analyses, presented in Chapter 4, that completeness of follow-up was related to characteristics at the time of stroke, such as age and stroke severity. Participants were also more likely to drop out of the study as they neared the end of their life, and deterioration was observed in health status during the follow-ups immediately prior to death and prior to dropout in those who survived. It is therefore highly likely that the missing data in the SLSR depends, at least in part, on current health status and may therefore be missing not at random (MNAR).

Many analyses of SLSR data focus on estimating prevalence of poor outcomes, of risk factors or proportions of patients accessing services. Although the missing data appear to be related to health status after stroke, particularly level of disability as measured by the barthel index, the distribution of Barthel scores is highly skewed, with the majority of participants scoring highly, or being classed as having mild or no disability. The impact that missing data had on rates estimated cross-sectionally

at individual follow-up points was found to be minimal, as those at greatest risk of drop-out represent a minority of participants. The impact was explored in a simulation study for which the results were summarised in Chapter 6. In this study a number of methods for handling missing data were explored in four scenarios, each of which made different assumptions about the mechanism underlying missing outcome data in the SLSR. When missing data were dependent on current level of disability at the time of follow-up, a scenario which was plausible based on exploratory analysis of the SLSR, prevalence rates of moderate-severe disability, inactivity and depression were underestimated using available case analysis, the approach used in the majority of SLSR papers published to date. Overall multiple imputation produced the least biased estimates of prevalence in this scenario. The bias was largest for inactivity, where differences between the observed and true prevalence rate of 50% were up to 5% points. Across the other three outcomes, where the true prevalence rate was lower, the maximum bias associated with multiple imputation was 2% points. Although available case analysis did underestimate true prevalence, it was only slightly worse than multiple imputation. Under the MNAR assumption, disability was underestimated by 5.2% and inactivity by 7.1% points. For depression, the difference between the true and estimates prevalence was less than 3% points.

The dangers of using single imputation methods were highlighted with substantial biases observed, even when the missing data were missing completely at random (MCAR). In the most extreme example the prevalence of inactivity was overestimated by almost 25% points following mode imputation. The other single imputation method applied in the simulation study was last observation carried forward, which in most cases produced unbiased prevalence estimates, though as each missing observation was imputed with a single value, the overall variability in the study is likely to be underestimated.

The hypothesis being tested in the thesis was that 'the application of suitable ana-

lytical methods to a longitudinal study with high levels of missing data will result in improved estimation when compared to available case analysis' and in terms of the analysis of rates post stroke the simulation study supported the hypothesis. The use of multiple imputation, and to a lesser extent inverse probability weighting, reduced bias observed when only available data were used. However, in the SLSR available case analyses are unlikely to result in large biases and are substantially less biased than using single imputation techniques. The lack of bias observed in the SLSR results from the distribution of the outcome data in the population of stroke survivors. Other data are collected in the SLSR at follow-up and there may potentially be other outcomes which are associated with missingness. To explore the potential impact of incomplete data on other outcomes, or in other studies, it is therefore important when exploring the data to consider not just the relationship between the observed data and missingness, but also the distribution of the variable(s) on which it is hypothesised that missing data might depend upon.

The measures of poor outcomes used in the simulation study were all derived from underlying scales. Imputation methods were applied directly to the categorised outcomes and also to the original scales with no substantial differences observed.

The impact of missing follow-up data on associations between baseline characteristics and outcomes after stroke was also assessed, with results presented in chapter 7. A number of models were fitted to the SLSR data which each made different assumptions about the missing data mechanism. There was no evidence of any difference in the regression coefficients from models which allowed for a missing not at (MNAR) mechanism and standard longitudinal models which assumed the data were MCAR or missing at random (MAR). This suggests that the relationship between the baseline characteristics and outcomes studied in this thesis may not differ in those with and without complete data. This finding is discussed further in Section 8.5.

In the following sections of this chapter the results described above are discussed in more detail, with consideration given to the relevance to other studies and implications for future work.

8.2 Incomplete follow-up in the SLSR and other cohort studies

One of the objectives of the thesis was to describe the patterns and predictors of missing data in the South London Stroke Register. Across the three month and annual follow-ups up to five years after stroke up to 30% of survivors did not complete the scheduled follow-up. Among those, between half and three quarters were classed as having dropped out and they did not complete any further follow-ups until the end of the study or death. Other participants did return to the study and completed a follow-up at at least one time point. There were a variety of reasons for the data being missing; some participants were registered too late after stroke to participate in the earlier follow-ups, some refused to participate in current follow-up or in all future follow-ups, and others were not contactable.

At any given follow-up point younger participants with less severe strokes were the most likely to not complete follow-up. Looking at the association between dropout and outcomes after stroke revealed that participants who dropped out earliest were more disabled and less active across all follow-ups and outcomes appeared to worsen sharply between the follow-ups immediately prior to dropout.

The missing data in the SLSR is not atypical of that found in other cohort studies. The North East Melbourne Stroke Incidence study (MENISIS) is a population based stroke study based in Melbourne Australia [8]. At three months after stroke they reported 63% of survivors were followed up, and 67% followed up at one year, a

similar rate as observed in the SLSR. At three months 28 of the 60 survivors did not complete follow-up as they were referred to the study too late [8]. At two and five years after stroke 15% and then 20% were lost to follow-up [9, 10]. As in the SLSR, younger participants were overall less likely to complete follow-up, as were those born outside of Australia [10].

Some stroke studies do achieve higher follow-up rates. Across Europe, follow-up rates in a German population based stroke study were 82% at one year [13], 95% at four and 16 months in a Swedish population based study [14] and 93% at one year in Greece [15].

In Auckland, New Zealand, a population based study achieved a follow-up rate of 94% at six months after stroke [11] and 93% at 21 years [12]. In Perth, Australia, another population based cohort study achieved a follow-up rate of 96% of survivors recruited to the study between 1989 and 90. The rate was lower among participants recruited six years later, who had a five year follow-up rate of 84%.

The Copenhagen heart study followed a sample of the general population aged 18 years and over recruited from a well defined area in Copenhagen, Denmark [208]. Across four waves of data collection between 64% and 81% of those who invited to each wave took part. In this study, low income, being single, widowed or divorced, being in unskilled employment or self employment, living alone, being less well educated, high alcohol consumption, being a heavy smoker, inactivity and obesity were all associated with increasing odds of dropping out [208].

Another Danish study followed a cohort with coronary heart disease who had undergone a coronary angioplasty [27]. Data were collected at eight time points over the 36 months following the angioplasty. The cohort included 1726 participants and of the 1652 alive at the end of follow-up, only 470 completed all follow-ups. Partici-

pants who failed to respond to any follow-up, and those who initially responded but subsequently dropped out were significantly older, were more likely to be smokers, had more co-morbidities, more severe disease and had lower levels of education

Similar predictors of attrition including age, ill-health, low levels of education or low socio-economic status and being from a minority ethnic group have also been observed in non cardiovascular studies [17, 20–26, 28–31].

The decline in health status prior to death observed in the SLSR has also been observed in other cohort studies [32, 209, 210], as has the association between death soon after the follow-up point and non-response [211].

In summary, although there are some differences in rates of missing data between studies, missing data are a problem in most, if not all, longitudinal studies. Participant characteristics measured at baseline are often shown to be associated with missing data, with age, variables reflecting low socioeconomic status and ill health consistently shown to be predictors of incomplete follow-up, particularly in studies following elderly cohorts. While the analysis presented in this thesis looks specifically at missing data in the SLSR, the findings may be generalisable to other studies, particularly those assessing health status over time.

Ideally the best way to handle missing data is to avoid it in the first place. SLSR participants are contacted first by telephone and then by post or email if required. Very few participants return postal questionnaires without having first been contacted by phone. When not possible to contact a participant then medical records are checked to ensure there has not been a change of address or telephone. Despite this lost to follow-up rates remain relatively high and other studies have shown non-response to be increasing over time. In a household survey in the United States participants were contacted and completed surveys by telephone [212]. Response rates decreased

from 1979 to 1996 but the rate of decrease became significantly larger between 1996 and 2003. Increasing difficulty in contacting participants by telephone is likely to be driven, at least in part, by people being more likely to regularly change mobile numbers and to screen or block calls from unknown numbers with caller identification services being widely available.

Offering incentives, making repeated contact (e.g. by sending postal or telephone reminders to complete postal questionnaires) and offering a range of data collection methods have been shown to increase retention in a number of population based studies [213]. In the SLSR, in the majority of cases where the follow-up was not completed, it was not possible to contact the participant. Offering incentives is unlikely to be beneficial in most of these cases. The SLSR also already contacts participant in a number of ways and participants are able to complete follow-ups by postal questionnaire, over the telephone or by email where face to face interviews are not possible or an alternative method is preferred. As it seems much of the missing data are unavoidable, understanding the impact it has on analyses is therefore important.

8.3 Impact of missing data on estimates of prevalence of poor outcomes

A simulation study was used to compare methods for handling missing in the estimation of prevalence rates of poor outcome after stroke. Previous analysis of the SLSR data have largely relied on available case analysis [158–161, 165], with inverse probability weighting (IPW) being applied in a few recent studies [162]. In the simulation study, estimates from available case analyses only resulted in slightly larger biases than those from IPW, which in turn were only marginally larger than those from multiple imputation. In multiple imputation outcomes measured at other time points were used in the imputation models along with baseline data, whereas in

IPW only baseline data were considered.

Engels et al (2003) used data from the Cardiovascular Health Study, a population based study which aimed to identify predictors of coronary heart disease and stroke in adults aged 65 and older, to compare imputation methods for incomplete follow-up data [85]. Known values on scales measuring depression, weight, cognitive function and self rated health, which followed one or more missing follow-ups were identified and assigned as being missing. These were then imputed using number of different methods and the imputed values compared to the true value. Only single imputation methods were applied, but it was found that methods which used any follow-up data available from the participant were far superior to those which imputed using population values or using baseline data only.

The model for non-response in the IPW approach included baseline characteristics to estimate the probability of response at each of the two time points. Theoretically, this model could include outcomes at previous time points when missing data patterns are monotone. When there is also intermittent missingness, as in the SLSR, the ability of multiple imputation to be able to handle missing data in multiple variables is one of the major advantages it has over IPW.

Single imputation methods were also applied to the SLSR data, which in general performed very poorly. Mode, median, mean, regression and hotdeck imputations all produced very biased estimates of prevalence, with rates of disability being underestimated by up to 20% points and inactivity overestimated by up to 25% points. The highly biased point estimates also resulted in very low coverage for some of these methods. Last observation carried forward (LOCF) was the best of all the single observation methods, with bias similar to that observed in available case analysis. LOCF, despite still being widely applied, particularly in trials, can lead to very biased treatment effects and inflated type 1 error rates [89]. After stroke, health

status tends to remain relatively stable or gradually decline. Once a participant is classed as disabled, inactive, anxious or depressed, it is highly likely that they will have the same status at the next follow-up which explains why in this case LOCF produced largely unbiased estimates.

Available case analyses did not result in large biases in the SLSR when estimating rates of poor outcomes and appeared to provide adequate estimates of prevalence of the outcomes studied. As described above, this finding was largely due to the distribution of the variable on which missingness was assumed to depend in the simulations (and on which it it likely the missing data does depend based on exploratory analyses of the SLSR). Participants who had the lowest possible score on the Barthel index were assumed to be six times more likely to drop out as those with the highest possible score. However, Barthel index is skewed, and most participants have high scores, so those at greatest risk of drop out represented a minority. Had it been assumed that those who were most well were most likely to drop out, available case analyses would have given highly biased results. Therefore, when presented with longitudinal data with drop-out it is important to consider not just the factors associated with drop out but also the distribution of these factors within the population to help understand the likely impact of incomplete data.

While biases were low, available case analysis was not unbiased and it was shown, bias can be reduced through the use of multiple imputation or inverse probability weighting. As IPW relies on complete data in the variables used in the model of response, multiple imputation would in general be the more favourable method as follow-up data from other time points can easily be incorporated. Ideally available case analyses should be conducted along with multiple imputation as a sensitivity analysis.

8.4 Imputations before and after dichotomisation

In the simulation study imputation methods were applied to the underlying scale used to derive the categorical outcome variables and compared to imputations made directly on the categorical form of the variables. Overall there was very little difference between the two. The majority of the methods applied were single imputation techniques, which are not recommended for use. Using multiple imputation, there was very little difference in the bias, coverage and standard errors when applied directly to the binary outcome or when treated as continuous and imputed before dichotomisation.

In multiple imputation linear regression models were used to impute the underlying scales. However, each of the scales have an upper and lower bound and without any constraints it is possible that values which are implausible may be imputed. In the simulation study the scales were dichotomised and so no restrictions were placed on the range of imputed values. The majority of follow-up data in the SLSR is categorical by design, but if the original forms of these scales are to be used in analysis then the appropriateness of the imputation model would require further consideration.

8.5 Impact of missing data on identifying predictors of outcome

The final objective was to assess the impact of missing data when exploring associations between baseline characteristics of participants and post-stroke outcomes. The models that were fitted to the data assumed that the missingness was MCAR, MAR or MNAR and there were very few differences between the parameter estimates from the models, though the MNAR and MIGEE models had larger standard errors than the others. As described in Chapter 2 Section 2.2.5.7 models which allow for a MNAR mechanism must be interpreted with caution and are best places in the

context of a sensitivity analysis [41]. Any model makes a specific assumption about the missing data mechanism and so a single model cannot be used to account for all possible MNAR processes. However, fitting one or more models which allow for plausible MNAR mechanisms and comparing parameter estimates to the primary model, which most likely assumes a MAR process, can be useful in checking robustness of conclusions against possible MNAR processes.

In this thesis shared parameter and pattern mixture models were applied. In the pattern mixture model the effect of covariates on outcome were allowed to vary across groups defined by time of exit from the study. There was no evidence that any differences in effect existed.

In the shared parameter model, the variance of the random effects that were shared between the drop-out model and the models for the outcomes were significantly different from zero. Though any hypothesis test needs to be treated with caution, this does suggest that the likely presence of an unmeasured trait that is associated with both drop out and outcome. Despite this, and in line with the results from ten pattern mixture model, when this train was allowed for, there is no change in the conclusions drawn from the model in terms of the effect of baseline covariates on outcome.

A number of other studies have also found no evidence of any bias in the estimates of model coefficients in the presence of missing data. In a study which imputed food frequency information collected longitudinally, there was little difference on the estimated impact on survival using complete case or multiple imputation methods. Several imputation methods were applied, some of which attempted to take account of correlation in the food frequency data which was collected longitudinally, but the additional complexity of these methods did not appear to offer any benefit or produce different results compared with imputations which did not take this into

account and assumed a multivariate normal distribution [214]. A study of patients following coronary angioplasty over three years had similar dropout patterns to those observed in the SLSR [27]. Available case analysis and multiple imputations were conducted on the data from this study and simulations based on the data developed to explore the effect of a MNAR assumption. The outcome of interest here was SF-12 scores, a measure of quality of life, which is summarised on scale from 0-100. The analyses explored whether varying the assumptions regarding the missing data mechanism had any impact on the association between gender and quality of life. While the mean scores were slightly different before and after imputation, suggesting bias in the estimation of the mean, there was little difference in the estimated effect of gender on quality of life.

As described previously, low socioeconomic status is often cited as a predictor of non-response. Data from the Avon Longitudinal Study of Parents and Children has been used to study the effect of non-response on socio-economic inequalities. Non response was associated with socio-economic status and outcomes and while bias in the effect of inequalities increased as non-participation increased, the qualitative conclusions drawn from the data regarding inequalities were broadly similar even when less than half the original sample provide outcome data [215].

While in the SLSR and the other studies highlighted above it appears that in many cases the estimated association between baseline characteristics and outcome are not significantly influenced by missing follow-up data, even when it is likely MNAR, this is not always the case. In a study of depressive symptoms in an elderly cohort, similar trajectories to those observed in the SLSR were observed prior to dropout, i.e. outcomes were observed to worsened rapidly immediately prior to dropout [32]. A mixed model and two shared parameter models (one of which treated time to dropout as continuous and the other as discrete) which allowed for the very likely MNAR dropout process were applied to the data and across most of the variables

included the parameter estimates were consistent. However, the use of an anti-depressive drug was associated with a three-fold increase in odds of depression in the mixed model, but in the joint models the odds were increased sixteen fold.

Encrenaz et al (2005) also conducted analysis on a real dataset which included a drug use severity score in participants who has started maintenance therapy in their treatment for opiate addiction [216]. At the follow-up 18 months after recruitment only 38% of the cohort provided data. Using a random effects model both age and treatment setting were found to be associated with drug use. When a joint model was applied instead these associations diminished and were non-significant.

In both of these examples it was very likely that the missing data were strongly related to the outcome of interest. In the example in participants with depressive symptoms, it was only the variable that was also very strongly related to the outcome that was underestimated using a standard MAR model. In the SLSR, it appeared that current level of disability and activity level were most strongly associated with drop-out, while there was less evidence of a change in level of depression of anxiety prior to dropout. In the models for level of disability and inactivity after stroke, disability at the time of stroke was very strongly associated with the outcomes. Multinomial models were applied to the disability data and it was not possible to fit MNAR models. In the analysis of activity level using proportional odds models, shared parameter models were fitted and there was no evidence of any bias in the estimated relationship between disability at the time of stroke and activity level.

Six covariates were included in the models applied to the SLSR data. While there was no evidence of bias across these six covariates it is possible the same might not be true of other factors recorded in the SLSR. Where the outcome of interest is likely to be strongly associated with the missing data mechanism then it seems particularly important to ensure models are not being influenced by the missing

data. Although no evidence of bias was found in the work presented in the thesis, there remains a need to ensure appropriate sensitivity analyses are carried out.

When conducting analysis the choice of a GEE (i.e. a population averaged) approach or a random effects model (i.e. a subject specific model) should be driven by the aim of the analysis [45]. There were no differences observed between GEEs in this study, which assume MCAR data, and weighted GEE (WGEE) or multiple imputation GEEs (MIGEE) which both allow for MAR data. MIGEE have been shown to be more robust against specification of the underpinning model and also have the advantage that missing covariates can be incorporated [60]. If a GEE approach is undertaken and the data are unlikely to be MCAR, to fit a model more robust against then missing data then MIGEE would be preferred over a standard GEE or WGEE.

Random effects models are valid under MAR assumptions without the need for imputation or weighting and applying multiple imputation prior to model fitting has been shown to have no benefit [146, 217, 218].

In addition to carrying out exploratory analyses of the data to help understand patterns and predictors of missing data an appropriate sensitivity analysis should also be carried out using MNAR models where possible. In this thesis pattern mixture and shared parameter random effects models were applied to the data. It is also possible to construct pattern mixture models for GEE based analyses and for continuous or binary data selection models may also be appropriate [41]. Although software and packages to fit these models are becoming increasingly available fitting the models can be very computationally intensive and models may fail to converge when there are multiple covariates. As described in Chapter 5, this was a particular issue in this thesis for the proportional odds and multinomial models. Where it is not possible to fit a MNAR model analogous to the main multivariable model,

univariable analyses could be conducted where possible.

8.6 Strengths and limitations

The work presented in this thesis has a number of strengths and limitations resulting from the design and scope of the included studies.

8.6.1 Simulated versus empirical data

A mix of simulations and analysis of a 'real' dataset were used in this thesis. Simulation studies allow properties of methods to be assessed by comparing results to a known 'true' value [219]. One of the major limitations of simulations studies is that they often lack the complexity observed in real datasets. To investigate the impact of missing data on estimates of prevalence of poor outcomes after stroke, real data from the SLSR was used with missing values added to the dataset using simulation methods. The distribution of, and relationship between, the baseline and outcome variables used therefore reflected a real scenario. The missing data were simulated so as to reflect rates of dropout and intermittent missingness in the SLSR as closely as possible. The mechanisms used to simulate the missing data were based on plausible scenarios following exploratory analysis of the SLSR, including one in which a MNAR assumption was made. As the true missing data mechanism in the SLSR is not known it's not possible to replicate exactly the true missing data mechanism. Simulating missing data according to four possible scenarios did however give an indication of the performance of methods and to estimate potential biases which may result from incomplete follow-ups.

In the study exploring the effect of missing data on associations between baseline data and outcome after stroke models were applied directly to the full SLSR dataset. In theory, had the models been applied to the simulation datasets created in the first study then parameter estimates could have also been compared to a known true

value. This was not feasible, as many of the models applied to the data took up to a day to run individually on a personal PC. In the simulation study 1000 datasets were used in each of four scenarios, and so attempting to fit the models in every one of these would not be possible without significantly more powerful computing resources. Applying the models directly to the full dataset did still have an advantage over a simulation study. The mechanism and patterns of missing data were those actually observed and no assumptions were required to attempt to produce realistic missing data mechanisms as in the simulation study.

8.6.2 Item non-response

All the work presented focused on missing outcome data which is arguably the source of missing data most likely to result in bias. In the SLSR some items are not completed at follow-up or in the baseline interview. For the scales used in this thesis, in some cases the full scale was not completed and in others single items were missing and therefore the full scale not computed, although this was rare. Overall, item non-response, with the exception of the hospital anxiety and depression scale (HADS) was low. In the work presented in this thesis missing data in the outcomes were treated the same regardless of the reason for being missing. Characteristics which predicted missing HADS in participants who had participated in a follow-up were similar to those which predicted non-participation in the follow-up. It is possible that HADS and the association with baseline characteristics may be different in the those missing only this scale.

Treating item non-response and non-participation in a follow-up in the same way is common. A review of 262 studies with missing data published in *Epidemiology*, the *International Journal of Epidemiology* and the *Americal Journal of Epidemiology* revealed that in almost half (46%) it was not possible to distinguish between item non-response and non-participation [7]. In many cases the underlying mechanism may be quite different for the two types of missing data. Where there is substantial

item non-response at follow-up care needs to be taken where those participants may differ from those who did not complete the follow-up at all.

Six variables representing baseline characteristics of the SLSR participants were used in the simulation studies and included in the models. These variables were selected as they are known to be associated with poor outcome after stroke, completeness of follow-up or both, and had very little missing information. Other variables collected at baseline are less complete. The majority of SLSR analyses thus far have used an available case analysis, with participants with missing baseline or follow-up data excluded. When fitting models incorporating baseline data multiple imputation can reduce bias in covariates when data are MCAR or MAR, though often available case is not more biased [220]. When data are MNAR it has been shown that parameter estimates can be biased [221]. The impact of missing baseline data in the SLSR is an area which could be explored further.

8.6.3 Handling missing data due to death

The focus of the thesis was on handling potentially MNAR outcome data in the SLSR. By far the largest source of missing data in the SLSR is due to non-response at follow-up.

In the SLSR mortality rates are high, with over 50% of the original cohort having died by five years after stroke. In the simulation study the population of interest was survivors and so the was no bias as a result of missing data due to death. In the study of the association between baseline characteristics and outcome, in all but the weighted GEEs, missing due to death was in no way accounted for. However, in the WGEE two approaches were taken - a mortal analysis and an immortal analysis. In the former any participant who was missing due to death was excluded from the point onwards, where as in the later those who were missing due to death were treated the same as those missing due to dropout. There were no differences

observed between these two models suggesting missingness due to death may have little impact.

Some studies attempt to address survivor biases which may arise when participants who died are excluded. Some simple approaches include imputing scores in the same way as missing data in survivors or assigning the worst possible outcome at all time points following death [27]. Kurland et al (2009) summarised the application of a number of models to simulated and real longitudinal studies with missing data due to death. They describe models such as a joint model for the outcome and death which can be used to estimate, for example, the probability of being alive and healthy at a given time point. GEEs fitted to the observed data describe the outcome in the cohort who are alive at each time point. Random effects models can be used, for example to fit unconditional models describing outcomes which are independent of death, or to model 'terminal decline' where the outcome is described in terms of time before death. Pattern mixture models with participants stratified by time of death can be useful for summarising individual trajectories when outcomes are related to time of death. Each of these models would not necessarily be appropriate in all settings, but instead the choice of model should be driven by the research question [222].

Applying these models in situations where the dropout is potentially MNAR and intermittent missingness also exist is not straight forward. A pattern mixture model which divides participants according to time of death and dropout may become impossible to fit when the data are divided into many groups.

A two stage imputation approach has also been suggested and demonstrated in longitudinal studies in elderly populations where there is a decline in the outcome of interest prior to death [223]. In this approach time until death is included as a predictor in the imputation model for the outcome. As, in many studies, some participants have unknown time of death, as they have not yet died, a two stage

procedure allows for missing data in the outcome and time until death to be treated separately. Two imputation models are then employed, the first describes the population response based on baseline covariates and then the second describes the response conditional on the time to dropout. Parameter estimates from the two models can then be combined.

In studies such as the SLSR, and others in elderly or sick cohorts, death is often an outcome of interest and so it is arguable that data missing due to death should never be thought of as missing data [224]. Further, in cohort studies where the sample are an unbiased subset of the population of interest at baseline, any deaths and predictors of deaths which occur in the sample will also occur in the population. Therefore it is likely that bias from dropout, giving an unrepresentative sample would be greater than biases due to death [225].

Knowledge of time of death may however be useful in reducing biases associated with summarising outcomes after stroke. In the simulation study reported in this thesis, missing data in the third scenario were simulated by allowing missingness to depend on time of death as well as baseline characteristics. However, the methods for handling missing data used did not take into account the time of death. This reflects normal practice where in the estimation of a rate at a given time point adjustment is not usually made for a future even (i.e. death). As drop-out and outcome are both observed to be associated with time of death in the SLSR, the incorporation of an indicator for time of death may help to minimise bias. The most appropriate way of including death in any analyses and any associated bias reduction is an area that warrants further investigation.

8.6.4 Implementation of missing data methods

A range of missing data methods were applied to the simulated and real SLSR data but some limitations exist due to the assumptions made regarding the rela-

tionship between observed data and missingness. In both the simulation study and in the longitudinal models applied to the whole SLSR dataset a subset of baseline covariates were used to simulate missing data patterns and in analyses. In the simulation study this meant that all factors associated with missing data were known and could be adjusted for. Inverse probability weighting relies on the assumption that the model for non-response is correctly specified [70] and as expected when all variables used to simulate the missing data were included as predictors in the model for non-response the estimates of prevalence were unbiased. In reality missing data in the SLSR are likely to depend on more factors than those included in the study. For example, social class is often cited as being associated with non-response. In the SLSR a substantial proportion of participants have unknown social class and so it was not included as a determinant of non-response in the simulated data. As well as ensuring all predictors of non-response are included, it is also important to ensure the model fits well and that continuous variables are handled appropriately. Age and GCS were both slightly skewed but did not appear to effect model fit and so were not transformed in any analyses. When including continuous covariates transformations may be necessary. Seaman et al provide a comprehensive guide to implementing inverse probability weighting, including the selection of variables for inclusion in the model of response and checking model fit [70].

Estimates in this thesis based on IPW did not differ substantially from available case analysis. The weights derived using baseline characteristic were not hugely variable. Including information from previous follow-ups may result in missingness models which are better able to discriminate between drop-outs and non-drop-outs whether the likelihood of drop-out is related to current and previous values of the outcome. However, where there is intermittent missingness the construction of appropriate models is challenging.

When IPW was used in conjunction with GEEs the intermittent missing values

were imputed using an average of the last and next completed observations. This was done to create a dataset with monotone missingness which is required for the construction of weights and the fitting of weighted GEEs. Although intermittent missingness appears less likely to be dependent on the outcome than drop-out, such an imputation approach assumes the data is not MAR and results in underestimated standard errors.

The implementation of the IPW approach for GEEs used in this thesis could be improved upon, perhaps by using multiple imputation for the intermittent missingness and then constructing weights based on the imputed values of outcome at earlier time points. However, there such an approach is unlikely to offer any advantage over MIGEE in which all the missing data are imputed using multiple imputation.

Multiple imputation also requires that the imputation model is correctly specified. The model should include all factors which are associated with outcome and non-response and should include all variables to be included in, and maintain the structure of, the analysis model [98]. As with inverse probability weighting in the simulation study, all variables associated with dropout were known, and so could be incorporated to obtain valid estimates under a MAR assumption.

In the imputation models used for both the simulation study and the MIGEEs, only outcomes prior to the one being imputed were included in the models. This was to avoid imputed values at early follow-ups being dependent on the imputed values at later follow-ups where a substantial proportion of participants had actually died. The imputations were also carried out separately for each outcome. The standard errors obtained following multiple imputation were only marginally lower than those using available cases. Further refinement of the imputation model may help to produce models which are better able to predict the posterior distribution of the missing values and in turn improve precision. The work in this thesis suggests that

MI is useful for reducing bias in the SLSR but further investigation is warranted to explore the impact of the structure of the imputation models. In particular, although only the first five years after stroke were considered in this thesis, some participants have completed over twenty follow-ups. When analysing long term outcomes it would not be feasible to construct models which incorporate all follow-up information. Further simulations could therefore be useful in providing guidance on choosing the most appropriate set of predictor variables which balance information on other outcome measures and the outcome of interest measured at other time points.

Both inverse probability weighting and multiple imputation were used in conjunction with generalised estimating equations in analysis of the SLSR data. It is likely in this dataset that other factors, in addition to the six included in the models, exist which predict non-response, or outcomes of interest. These additional variables may also have missing data themselves. Mice allows for missing data in covariates to be imputed at the same time as outcomes making it more attractive than inverse probability weighting which requires variables in the model of response to be complete. As a result of this it was also possible to include outcomes measured at previous follow-ups in imputation models. Non response in the SLSR was related to health status at follow-up and so incorporating this information resulted in lower biases when the data were missing not at random and multiple imputation was used.

Pattern mixture and shared parameter models were also applied to the SLSR data and allowed for the possibility of non-random dropout. In both cases the cohort were defined according to the time at which they left the study, be it due to death or dropout. In exploratory analyses of the SLSR deterioration were observed prior to death and prior to dropout in survivors, but at a slower rate than in those who died. Ideally patterns would be formed to allow for differences between those who dropped out and those who died. Increasing the number of groups or patterns in a

pattern mixture model was not feasible. The number of parameters required meant that analyses were restricted to univariable associations.

8.6.5 Alternative missing data methods

Many of the most common methods for handling missing data were included in the studies presented in this thesis, but they do not provide an exhaustive list. Handling missing data is an area of ongoing interest with new methods or developments and improvement of existing methods regularly being proposed. Multiple imputation produced the least biased estimates of the prevalence of poor outcomes after stroke and alongside available case analysis would be the method recommended for use in studies summarising outcomes after stroke.

Multiple imputation required that data are MAR but in the SLSR it is plausible that the data are MNAR. Fraser and Yan (2007) illustrated that valid estimates can be obtained under MNAR mechanisms by obtaining data from participants who were originally missing [226]. By re-contacting a random sub-sample of participants with missing data it was shown that obtaining missing information from as few as 10% of those missing and using the new information to update the imputation model provides estimates which are unbiased in MNAR situations. While this may be possible in some situations, in the many cohort studies like the SLSR, every effort is put into contacting the participant to prevent missing data in the first place. It would therefore be extremely difficult to contact those who were missing. Even if it was possible to obtain data from 10% of the original non-responders it is unlikely that they would be a random subsample of all non-responders, and instead would likely be more similar to the original responders.

Another extension of multiple imputation involves imputations being combined with inverse probability weights. As describe previously multiple imputation requires on the correct specification of the imputation model while inverse probability weighting

relies on correct specification of the model of response. Both methods can be sensitive to mis-specification of the respective model. In a doubly robust approach the methods are combined by including only participants who, for example attended a follow-up. Multiple imputation is then carried out for these participants and within the imputed dataset analyses are weighted to account for participants not included due to non-response [71].

When analysis aims to identify predictors of outcome by modelling the longitudinal data, sensitivity analyses should be conducted after fitting standard longitudinal models. Pattern mixture models can be sensitive to model mis-specification [123]. Demitris (2005) proposed an alternative pattern mixture model which employs Bayesian techniques to smooth parameter estimates across patterns [227]. This means that information from within other patterns or groups is also used when estimating parameters within a group [227] and through the use of simulations and analysis of empirical data it has been shown the model is more robust against model misspecification than conventional pattern mixture models [227, 228].

8.7 Recommendations for handling missing data

Until recently analysis of SLSR data has been via available case analyses. Most studies did report rates of missing data and compared the characteristics of those with and without missing data but the impact that any differences between those with and without missing data may have on findings had not been investigated. Many researchers have and continue to use the SLSR data without any guidance on how best to deal with missing data. Similarly many analyses of other cohort studies are published, which describe missing data and factors associated with non-response, but which present findings from available case analyses only. Many of these studies exhibit patterns and predictors of non-response which are similar to those in the SLSR and so recommendations for handling missing data in the SLSR may also apply elsewhere.

Based on the findings presented in this thesis missing data in the SLSR is unlikely to result in substantial biases when summarising outcomes after stroke using available case analyses as even though the missingness is likely dependent on health status, the majority of participants are relatively healthy. Despite this, the use of appropriate sensitivity analyses should always be carried out to ensure findings from individual studies are as robust as possible [229–231]. For future studies of outcome after stroke it is recommended that

1. In studies summarising data recorded at follow-up multiple imputation should be applied in addition to available case analysis to ensure minimal bias.

2. The multiple imputation should be performed using chained equations which make use of outcomes recorded at other time points.

3. In studies measuring associations between baseline data and follow-up standard longitudinal models (GEE or random effects models) should be followed by a sensitivity analysis applying an appropriate MNAR model (i.e. a shared parameter or pattern mixture model) where possible.

8.8 Future research

The work presented in this thesis focuses on four outcomes; disability, inactivity, anxiety and depression. The SLSR collects data on many more outcomes and measures of resource use after stroke. Apart from the scales described and applied in this thesis most of the data are recorded using binary, ordinal or nominal variables. As the outcomes describe in this thesis are most often categorised for analysis (despite this resulting in a loss of information) the categorical or binary formats of these scales were used. Therefore findings may also apply to other outcomes not considered here. It was shown that biases associated with prevalence estimates are not likely to be substantial even when missing data were related strongly to current

(poor) health status given that the majority of participants are in relatively good health. However, as a simulation study was used to estimate bias, the true bias resulting from missed follow-ups cannot be accurately determined.

Work is ongoing to link the SLSR with a number of other databases, including Lambeth datanet. This database includes GP records from the borough of Lambeth, one of the two boroughs covered by the SLSR. In the first merging of the two datasets over 800 SLSR participants were matched to data in Lambeth datanet. Using participants who appear in both datasets it may be possible to design a study to compare factors such as the prevalence of risk factors, and use of services in participants who did and did not drop out of the SLSR.

Similar study designs have been used in other cohorts. In a study of subjects with gastrointestinal disorders (GI), a response rate to a postal survey of 52% was achieved [232]. Medical records were then reviewed for a random sample of responders and non-responders which revealed that non-response was not associated with GI symptoms or specific diagnoses, the outcomes of interest in the survey. Comparisons have also been made between risk factor profiles and service use in elderly population based cohort people aged 50+ and data from primary care consultations [233]. In this case the data were not directly linked but comparisons between the two data sets suggested that there was no increase in bias in the estimation of the prevalence of risk factors resulting from dropout in the cohort study. Work is ongoing to merge the SLSR and Lambeth datanet databases and produce a dataset in which it would be feasible to conduct such analyses.

References

[1] M. Szklo, "Population-based cohort studies," *Epidemiologic reviews*, vol. 20, no. 1, p. 81, 1998.

[2] M. Chatfield, C. Brayne, and F. Matthews, "A systematic literature review of attrition between waves in longitudinal studies in the elderly shows a consistent pattern of dropout between differing studies," *Journal of clinical epidemiology*, vol. 58, no. 1, pp. 13–19, 2005.

[3] A. Karahalios, L. Baglietto, J. B. Carlin, D. R. English, and J. A. Simpson, "A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures," *BMC medical research methodology*, vol. 12, no. 1, p. 96, 2012.

[4] K. M. Jamsen, J. Ilomäki, S. N. Hilmer, N. Jokanovic, E. C. Tan, and J. S. Bell, "A systematic review of the statistical methods in prospective cohort studies investigating the effect of medications on cognition in older people," *Research in Social and Administrative Pharmacy*, 2015.

[5] K. L. Masconi, T. E. Matsha, J. B. Echouffo-Tcheugui, R. T. Erasmus, and A. P. Kengne, "Reporting and handling of missing data in predictive research for prevalent undiagnosed type 2 diabetes mellitus: a systematic review," *EPMA Journal*, vol. 6, no. 1, p. 7, 2015.

[6] M. L. Bell, M. Fiero, N. J. Horton, and C.-H. Hsu, "Handling missing data in rcts; a review of the top medical journals," *BMC medical research methodology*, vol. 14, no. 1, p. 118, 2014.

[7] I. Eekhout, R. M. de Boer, J. W. Twisk, H. C. de Vet, and M. W. Heymans, "Missing data: a systematic review of how they are reported and handled," *Epidemiology*, vol. 23, no. 5, pp. 729–732, 2012.

[8] H. M. Dewey, J. Sturm, G. A. Donnan, R. A. Macdonell, J. J. McNeil, and A. G. Thrift, "Incidence and outcome of subtypes of ischaemic stroke: initial results from the north east melbourne stroke incidence study (nemesis)," *Cerebrovascular Diseases*, vol. 15, no. 1-2, pp. 133–139, 2003.

[9] J. W. Sturm, G. A. Donnan, H. M. Dewey, R. A. Macdonell, A. K. Gilligan, V. Srikanth, and A. G. Thrift, "Quality of life after stroke the north east melbourne stroke incidence study (nemesis)," *Stroke*, vol. 35, no. 10, pp. 2340–2345, 2004.

[10] S. L. Paul, J. W. Sturm, H. M. Dewey, G. A. Donnan, R. A. Macdonell, and A. G. Thrift, "Long-term outcome in the north east melbourne stroke incidence study predictors of quality of life at 5 years after stroke," *Stroke*, vol. 36, no. 10, pp. 2082–2086, 2005.

[11] N. Kerse, V. Parag, V. L. Feigin, H. McNaughton, M. L. Hackett, D. A. Bennett, C. S. Anderson, *et al.*, "Falls after stroke results from the auckland regional community stroke (arcos) study, 2002 to 2003," *Stroke*, vol. 39, no. 6, pp. 1890–1893, 2008.

[12] C. S. Anderson, K. N. Carter, W. J. Brownlee, M. L. Hackett, J. B. Broad, and R. Bonita, "Very long-term outcome after stroke in auckland, new zealand," *Stroke*, vol. 35, no. 8, pp. 1920–1924, 2004.

[13] P. L. Kolominsky-Rabas, M.-J. Hilz, B. Neundoerfer, and P. U. Heuschmann, "Impact of urinary incontinence after stroke: Results from a prospective

population-based stroke register," *Neurourology and urodynamics*, vol. 22, no. 4, pp. 322–327, 2003.

[14] A.-C. Jönsson, I. Lindgren, B. Hallström, B. Norrving, and A. Lindgren, "Prevalence and intensity of pain after stroke: a population based study focusing on patients perspectives," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 77, no. 5, pp. 590–595, 2006.

[15] K. Vemmos, M. Bots, P. Tsibouris, V. Zis, C. Takis, D. Grobbee, and S. Stamatelopoulos, "Prognosis of stroke in the south of greece: 1 year mortality, functional outcome and its determinants: the arcadia stroke registry," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 69, no. 5, pp. 595–600, 2000.

[16] L. C. Thygesen, C. Johansen, N. Keiding, E. Giovannucci, and M. Grønbæk, "Effects of sample attrition in a longitudinal study of the association between alcohol intake and all-cause mortality," *Addiction*, vol. 103, no. 7, pp. 1149–1159, 2008.

[17] K. Gustavson, T. von Soest, E. Karevold, and E. Røysamb, "Attrition and generalizability in longitudinal studies: findings from a 15-year population-based study and a monte carlo simulation study," *BMC Public Health*, vol. 12, no. 1, p. 918, 2012.

[18] M. A. Badawi, W. W. Eaton, J. Myllyluoma, L. Weimer, and J. Gallo, "Psychopathology and attrition in the baltimore eca 15-year follow-up 1981–1996," *Social psychiatry and psychiatric epidemiology*, vol. 34, no. 2, pp. 91–98, 1999.

[19] A. J. M. Van Loon, M. Tijhuis, H. S. J. Picavet, P. G. Surtees, and J. Ormel, "Survey non-response in the netherlands: effects on prevalence estimates and associations," *Annals of epidemiology*, vol. 13, no. 2, pp. 105–110, 2003.

[20] S. Demarest, J. Van der Heyden, R. Charafeddine, J. Tafforeau, H. Van Oyen, and G. Van Hal, "Socio-economic differences in participation of households

in a belgian national health survey," *The European Journal of Public Health*, p. cks158, 2012.

[21] M. Lindén-Boström and C. Persson, "A selective follow-up study on a public health survey," *The European Journal of Public Health*, p. ckr193, 2012.

[22] M. Goldberg, J. F. Chastang, M. Zins, I. Niedhammer, and A. Leclerc, "Health problems were the strongest predictors of attrition during follow-up of the gazel cohort," *Journal of clinical epidemiology*, vol. 59, no. 11, pp. 1213–1221, 2006.

[23] A. Boys, J. Marsden, G. Stillwell, K. Hatchings, P. Griffiths, and M. Farrell, "Minimizing respondent attrition in longitudinal research: practical implications from a cohort study of adolescent drinking," *Journal of adolescence*, vol. 26, no. 3, pp. 363–373, 2003.

[24] M. Garcia, E. Fernandez, A. Schiaffino, C. Borrell, M. Marti, J. M. Borras, *et al.*, "Attrition in a population-based cohort eight years after baseline interview: The cornella health interview survey follow-up (chis. fu) study," *Annals of epidemiology*, vol. 15, no. 2, pp. 98–104, 2005.

[25] J. W. Krellman, S. A. Kolakowsky-Hayner, L. Spielman, M. Dijkers, F. M. Hammond, J. Bogner, T. Hart, J. B. Cantor, and T. Tsaousides, "Predictors of follow-up completeness in longitudinal research on traumatic brain injury: Findings from the national institute on disability and rehabilitation research traumatic brain injury model systems program," *Archives of physical medicine and rehabilitation*, vol. 95, no. 4, pp. 633–641, 2014.

[26] R. J. Lacey, K. P. Jordan, and P. R. Croft, "Does attrition during follow-up of a population cohort study inevitably lead to biased estimates of health status?," *PloS one*, vol. 8, no. 12, p. e83948, 2013.

[27] K. Biering, N. H. Hjollund, and M. Frydenberg, "Using multiple imputation to deal with missing data and attrition in longitudinal studies with repeated

measures of patient-reported outcomes," *Clinical epidemiology*, vol. 7, p. 91, 2015.

[28] C. Lee, A. J. Dobson, W. J. Brown, L. Bryson, J. Byles, P. Warner-Smith, and A. F. Young, "Cohort profile: the australian longitudinal study on women's health," *International Journal of Epidemiology*, vol. 34, no. 5, pp. 987–991, 2005.

[29] W. J. Van Der Veen, K. Van Der Meer, and B. W. Penninx, "Screening for depression and anxiety: correlates of non-response and cohort attrition in the netherlands study of depression and anxiety (nesda)," *International journal of methods in psychiatric research*, vol. 18, no. 4, pp. 229–239, 2009.

[30] S. Vega, J. Benito-León, F. Bermejo-Pareja, M. J. Medrano, L. M. Vega-Valderrama, C. Rodríguez, and E. D. Louis, "Several factors influenced attrition in a population-based elderly cohort: neurological disorders in central spain study," *Journal of clinical epidemiology*, vol. 63, no. 2, pp. 215–222, 2010.

[31] A. F. Young, J. R. Powers, and S. L. Bell, "Attrition in longitudinal studies: who do you lose?," *Australian and New Zealand journal of public health*, vol. 30, no. 4, pp. 353–361, 2006.

[32] C.-C. H. Chang, H.-C. Yang, G. Tang, and M. Ganguli, "Minimizing attrition bias: a longitudinal study of depressive symptoms in an elderly cohort," *International Psychogeriatrics*, vol. 21, no. 05, pp. 869–878, 2009.

[33] R. Little and D. Rubin, "Statistical analysis with missing data," *Hoboken (NJ): Wiley-Interscience*, 2002.

[34] A. Dempster, N. Laird, D. Rubin, *et al.*, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[35] J. Heckman, "Sample selection bias as a specification error," *Econometrica: Journal of the econometric society*, pp. 153–161, 1979.

[36] D. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, p. 581, 1976.

[37] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art.," *Psychological methods*, vol. 7, no. 2, p. 147, 2002.

[38] R. Little and D. Rubin, "Statistical analysis with missing data," 1987.

[39] P. Diggle, D. Farewell, and R. Henderson, "Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal," *Analysis*, vol. 56, no. Part 4, pp. 1–31, 2007.

[40] G. Fitzmaurice, N. Laird, and J. Ware, *Applied longitudinal analysis*. Wiley-IEEE, 2004.

[41] G. Molenberghs and M. Kenward, *Missing data in clinical studies*. John Wiley & Sons Inc, 2007.

[42] D. G. Altman and P. Royston, "The cost of dichotomising continuous variables," *Bmj*, vol. 332, no. 7549, p. 1080, 2006.

[43] J. Cohen, "The cost of dichotomization," *Applied Psychological Measurement*, vol. 7, no. 3, pp. 249–253, 1983.

[44] R. C. MacCallum, S. Zhang, K. J. Preacher, and D. D. Rucker, "On the practice of dichotomization of quantitative variables.," *Psychological methods*, vol. 7, no. 1, p. 19, 2002.

[45] P. Diggle, *Analysis of longitudinal data*. Oxford University Press, USA, 2002.

[46] N. Goldfarb, *An introduction to longitudinal statistical analysis: the method of repeated observations from a fixed sample*. Free Press, 1960.

[47] I. Carrière and J. Bouyer, "Choosing marginal or random-effects models for longitudinal binary responses: application to self-reported disability among older persons," *BMC Medical Research Methodology*, vol. 2, no. 1, p. 15, 2002.

[48] B. Everitt and D. Howell, *Encyclopedia of statistics in behavioral science*. John Wiley & Sons, 2005.

[49] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 42, no. 2, pp. 109–142, 1980.

[50] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, vol. 398. John Wiley & Sons, 2013.

[51] A. Agresti and M. Kateri, *Categorical data analysis*. Springer, 2011.

[52] R. Williams, "Generalized ordered logit/partial proportional odds models for ordinal dependent variables," *Stata Journal*, vol. 6, no. 1, pp. 58–82, 2006.

[53] J. S. Long and J. Freese, *Regression models for categorical dependent variables using Stata*. Stata press, 2006.

[54] G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, *Longitudinal data analysis*. CRC Press, 2008.

[55] F. B. Hu, J. Goldberg, D. Hedeker, B. R. Flay, and M. A. Pentz, "Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes," *American Journal of Epidemiology*, vol. 147, no. 7, pp. 694–703, 1998.

[56] J. I. Galbraith, S. L. Zeger, K.-Y. Liang, and P. S. Albert, "The interpretation of a regression coefficient," *Biometrics*, pp. 1593–1596, 1991.

[57] S. Zeger, K. Liang, and P. Albert, "The interpretation of a regression coefficient-response," 1991.

[58] S. L. Zeger, K.-Y. Liang, and P. S. Albert, "Models for longitudinal data: a generalized estimating equation approach," *Biometrics*, pp. 1049–1060, 1988.

[59] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," *Biometrics*, pp. 963–974, 1982.

[60] C. Beunckens, C. Sotto, and G. Molenberghs, "A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data," *Computational Statistics & Data Analysis*, vol. 52, no. 3, pp. 1533–1548, 2008.

[61] S. Zeger and K. Liang, "Longitudinal data analysis for discrete and continuous outcomes," *Biometrics*, vol. 42, no. 1, pp. 121–130, 1986.

[62] Y. Lee, J. A. Nelder, *et al.*, "Conditional and marginal models: another view," *Statistical Science*, vol. 19, no. 2, pp. 219–238, 2004.

[63] C. Szmaragd, P. Clarke, and F. Steele, "Subject specific and population average models for binary longitudinal data: a tutorial," *Longitudinal and Life Course Studies*, vol. 4, no. 2, pp. 147–165, 2013.

[64] R. Little, "Modeling the Drop-Out Mechanism in Repeated-Measures Studies.," *Journal of the American Statistical Association*, vol. 90, no. 431, 1995.

[65] J. Twisk and W. de Vente, "Attrition in longitudinal studies How to deal with missing data," *Journal of clinical epidemiology*, vol. 55, no. 4, pp. 329–337, 2002.

[66] A. N. Baraldi and C. K. Enders, "An introduction to modern missing data analyses," *Journal of School Psychology*, vol. 48, no. 1, pp. 5–37, 2010.

[67] R. M. Daniel, M. G. Kenward, S. N. Cousens, and B. L. De Stavola, "Using causal diagrams to guide analysis in missing data problems," *Statistical methods in medical research*, vol. 21, no. 3, pp. 243–256, 2012.

[68] J. Carpenter and M. Kenward, *Multiple imputation and its application*. John Wiley & Sons, 2012.

[69] C. Dufouil, C. Brayne, and D. Clayton, "Analysis of longitudinal studies with death and drop-out: a case study," *Statistics in medicine*, vol. 23, no. 14, pp. 2215–2226, 2004.

[70] S. R. Seaman and I. R. White, "Review of inverse probability weighting for dealing with missing data," *Statistical methods in medical research*, vol. 22, no. 3, pp. 278–295, 2013.

[71] S. R. Seaman, I. R. White, A. J. Copas, and L. Li, "Combining multiple imputation and inverse-probability weighting," *Biometrics*, vol. 68, no. 1, pp. 129–137, 2012.

[72] D. Clayton, D. Spiegelhalter, G. Dunn, and A. Pickles, "Analysis of longitudinal binary data from multi-phase sampling," *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pp. 71–87, 1998.

[73] J. M. Robins and R. D. Gill, "Non-response models for the analysis of non-monotone ignorable missing data," *Statistics in medicine*, vol. 16, no. 1, pp. 39–56, 1997.

[74] J. Carpenter, M. Kenward, and S. Vansteelandt, "A comparison of multiple imputation and inverse probability weighting for analyses with missing data," *Journal of the Royal Statistical Society, Series A*, vol. 169, no. 3, pp. 571–84, 2006.

[75] H. Bang and J. M. Robins, "Doubly robust estimation in missing data and causal inference models," *Biometrics*, vol. 61, no. 4, pp. 962–973, 2005.

[76] J. D. Kang and J. L. Schafer, "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," *Statistical science*, pp. 523–539, 2007.

[77] S. Vansteelandt, J. Carpenter, and M. G. Kenward, "Analysis of incomplete data using inverse probability weighting and doubly robust estimators," *Methodology*, 2010.

[78] T. A. Myers, "Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data," *Communication Methods and Measures*, vol. 5, no. 4, pp. 297–310, 2011.

[79] R. R. Andridge and R. J. Little, "A review of hot deck imputation for survey non-response," *International Statistical Review*, vol. 78, no. 1, pp. 40–64, 2010.

[80] P. Elliott and G. Hawthorne, "Imputing missing repeated measures data: how should we proceed?," *Australian and New Zealand Journal of Psychiatry*, vol. 39, no. 7, pp. 575–582, 2005.

[81] E. S. Nordholt, "Imputation: methods, simulation experiments and practical examples," *International Statistical Review*, vol. 66, no. 2, pp. 157–180, 1998.

[82] J. Siddique and T. R. Belin, "Multiple imputation using an iterative hot-deck with distance-based donor selection," *Statistics in medicine*, vol. 27, no. 1, pp. 83–102, 2008.

[83] S. M. Iacus and G. Porro, "Missing data imputation, matching and other applications of random recursive partitioning," *Computational statistics & data analysis*, vol. 52, no. 2, pp. 773–789, 2007.

[84] W.-Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[85] J. M. Engels and P. Diehr, "Imputation of missing longitudinal data: a comparison of methods," *Journal of clinical epidemiology*, vol. 56, no. 10, pp. 968–976, 2003.

[86] G. Kalton, "Compensating for missing survey data.," 1983.

[87] O. Siddiqui, H. J. Hung, and R. O'Neill, "Mmrm vs. locf: a comprehensive comparison based on simulation study and 25 nda datasets," *Journal of biopharmaceutical statistics*, vol. 19, no. 2, pp. 227–246, 2009.

[88] G. Molenberghs, H. Thijs, I. Jansen, C. Beunckens, M. Kenward, C. Mallinckrodt, and R. Carroll, "Analyzing incomplete longitudinal clinical trial data," *Biostatistics*, vol. 5, no. 3, p. 445, 2004.

[89] R. Cook, L. Zeng, and G. Yi, "Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation," *Biometrics*, vol. 60, no. 3, pp. 820–828, 2004.

[90] P. Lane, "Handling drop-out in longitudinal clinical trials: a comparison of the LOCF and MMRM approaches," *Pharmaceutical statistics*, vol. 7, no. 2, p. 93, 2008.

[91] G. Liu and A. L. Gould, "Comparison of alternative strategies for analysis of longitudinal trials with dropouts," *Journal of biopharmaceutical statistics*, vol. 12, no. 2, pp. 207–226, 2002.

[92] C. H. Mallinckrodt, C. J. Kaiser, J. G. Watkin, M. J. Detke, G. Molenberghs, and R. J. Carroll, "Type i error rates from likelihood-based repeated measures analyses of incomplete longitudinal data," *Pharmaceutical Statistics*, vol. 3, no. 3, pp. 171–186, 2004.

[93] L. Tang, J. Song, T. Belin, and J. Un
"utzer, "A comparison of imputation methods in a longitudinal randomized clinical trial," *Statistics in Medicine*, vol. 24, no. 14, pp. 2111–2128, 2005.

[94] M. Kenward and G. Molenberghs, "Last Observation Carried Forward: A Crystal Ball?," *Journal of Biopharmaceutical Statistics*, vol. 19, no. 5, pp. 872–888, 2009.

[95] D. Rubin, "Multiple imputation for nonresponse in surveys ," *NY: John Wiley & Sons*, 1987.

[96] D. B. Rubin, "Multiple imputation after 18+ years," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 473–489, 1996.

[97] J. A. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter, "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls," *Bmj*, vol. 338, p. b2393, 2009.

[98] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: issues and guidance for practice," *Statistics in medicine*, vol. 30, no. 4, pp. 377–399, 2011.

[99] J. W. Graham, A. E. Olchowski, and T. D. Gilreath, "How many imputations are really needed? some practical clarifications of multiple imputation theory," *Prevention Science*, vol. 8, no. 3, pp. 206–213, 2007.

[100] T. E. Bodner, "What improves with increased missing data imputations?," *Structural Equation Modeling*, vol. 15, no. 4, pp. 651–675, 2008.

[101] J. L. Schafer, *Analysis of incomplete multivariate data*. CRC press, 1997.

[102] K.-H. Li, T. E. Raghunathan, and D. B. Rubin, "Large-sample significance levels from multiply imputed data using moment-based statistics and an f reference distribution," *Journal of the American Statistical Association*, vol. 86, no. 416, pp. 1065–1073, 1991.

[103] A. R. T. Donders, G. J. van der Heijden, T. Stijnen, and K. G. Moons, "Review: a gentle introduction to imputation of missing values," *Journal of clinical epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006.

[104] M. Kenward and J. Carpenter, "Multiple imputation: current perspectives," *Statistical Methods in Medical Research*, vol. 16, no. 3, p. 199, 2007.

[105] X.-L. Meng, "Multiple-imputation inferences with uncongenial sources of input," *Statistical Science*, pp. 538–558, 1994.

[106] J. W. Graham, "Missing data analysis: Making it work in the real world," *Annual review of psychology*, vol. 60, pp. 549–576, 2009.

[107] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: what is it and how does it work?," *International journal of methods in psychiatric research*, vol. 20, no. 1, pp. 40–49, 2011.

[108] K. J. Lee, J. C. Galati, J. A. Simpson, and J. B. Carlin, "Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of non-linear effects in a large cohort study," *Statistics in medicine*, vol. 31, no. 30, pp. 4164–4174, 2012.

[109] I. Lipkovich, Z. Kadziola, L. Xu, T. Sugihara, and C. H. Mallinckrodt, "Comparison of several multiple imputation strategies for repeated measures analysis of clinical scales: to truncate or not to?," *Journal of biopharmaceutical statistics*, vol. 24, no. 4, pp. 924–943, 2014.

[110] N. J. Horton, S. R. Lipsitz, and M. Parzen, "A potential for bias when rounding in multiple imputation," *The American Statistician*, vol. 57, no. 4, pp. 229–232, 2003.

[111] C. A. Bernaards, T. R. Belin, and J. L. Schafer, "Robustness of a multivariate normal approximation for imputation of incomplete binary data," *Statistics in medicine*, vol. 26, no. 6, pp. 1368–1382, 2007.

[112] A.-F. Donneau, M. Mauer, G. Molenberghs, and A. Albert, "A simulation study comparing multiple imputation methods for incomplete longitudinal ordinal data," *Communications in Statistics-Simulation and Computation*, vol. 44, no. 5, pp. 1311–1338, 2015.

[113] A.-F. Donneau, M. Mauer, P. Lambert, G. Molenberghs, and A. Albert, "Simulation-based study comparing multiple imputation methods for non-monotone missing ordinal data in longitudinal settings," *Journal of biopharmaceutical statistics*, vol. 25, no. 3, pp. 570–601, 2015.

[114] S. Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of statistical software*, vol. 45, no. 3, 2011.

[115] J. Carpenter and H. Goldstein, "Multiple imputation in mlwin," *Multilevel Modelling Newsletter*, vol. 16, pp. 9–18, 2005.

[116] J. R. Carpenter, H. Goldstein, M. G. Kenward, *et al.*, "Realcom-impute software for multilevel multiple imputation with mixed response types," *Journal of Statistical Software*, vol. 45, no. 5, pp. 1–14, 2011.

[117] C. Welch, J. Bartlett, and I. Petersen, "Application of multiple imputation using the two-fold fully conditional specification algorithm in longitudinal clinical data," *The Stata journal*, vol. 14, no. 2, p. 418, 2014.

[118] C. Beunckens, G. Molenberghs, and M. G. Kenward, "Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials," *Clinical Trials*, vol. 2, no. 5, pp. 379–386, 2005.

[119] J. Robins, A. Rotnitzky, and L. Zhao, "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data," *Journal of the American Statistical Association*, pp. 106–121, 1995.

[120] A. B. Troxel, S. R. Lipsitz, and T. A. Brennan, "Weighted estimating equations with nonignorably missing response data," *Biometrics*, pp. 857–869, 1997.

[121] J. S. Preisser, K. K. Lohman, and P. J. Rathouz, "Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random," *Statistics in medicine*, vol. 21, no. 20, pp. 3035–3054, 2002.

[122] J. Ibrahim and G. Molenberghs, "Missing data methods in longitudinal studies: a review," *Test*, vol. 18, no. 1, pp. 1–43, 2009.

[123] H. Demirtas and J. L. Schafer, "On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out," *Statistics in medicine*, vol. 22, no. 16, pp. 2553–2575, 2003.

[124] D. Hedeker and R. D. Gibbons, "Application of random-effects pattern-mixture models for missing data in longitudinal studies.," *Psychological methods*, vol. 2, no. 1, p. 64, 1997.

[125] J. W. Hogan and N. M. Laird, "Mixture models for the joint distribution of repeated measures and event times," *Statistics in medicine*, vol. 16, no. 3, pp. 239–257, 1997.

[126] R. Little, "Pattern-mixture models for multivariate incomplete data," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 125–134, 1993.

[127] G. Molenberghs, G. Fitzmaurice, M. G. Kenward, A. Tsiatis, and G. Verbeke, *Handbook of Missing Data Methodology*. CRC Press, 2014.

[128] R. J. Little, "Modeling the drop-out mechanism in repeated-measures studies," *Journal of the American Statistical Association*, vol. 90, no. 431, pp. 1112–1121, 1995.

[129] M. G. Kenward, "Selection models for repeated measurements with non-random dropout: an illustration of sensitivity," *Statistics in medicine*, vol. 17, no. 23, pp. 2723–2732, 1998.

[130] G. Kemmler, M. Hummer, C. Widschwendter, and W. Fleischhacker, "Dropout rates in placebo-controlled and active-control clinical trials of antipsychotic drugs: a meta-analysis," *Archives of general psychiatry*, vol. 62, no. 12, p. 1305, 2005.

[131] W. Abraham and D. Russell, "Missing data: a review of current methods and applications in epidemiological research," *Current Opinion in Psychiatry*, vol. 17, no. 4, p. 315, 2004.

[132] J. Hogan, J. Roy, and C. Korkontzelou, "Handling drop-out in longitudinal studies," *Statistics in Medicine*, vol. 23, no. 9, pp. 1455–1497, 2004.

[133] D. Adamis, "Statistical methods for analysing longitudinal data in delirium studies," *International Review of Psychiatry*, vol. 21, no. 1, pp. 74–85, 2009.

[134] "Sas 6 for windows, sas institute, cary (1997),"

[135] StataCorp, "Stata statistical software: Release 9," vol. College Station, TX: StataCorp LP, 2005.

[136] "Spss for windows, rel. 11.0.1. 2001. chicago: Spss inc.,"

[137] "R lme archive." https://mirrors.ebi.ac.uk/CRAN/src/contrib/Archive/lme/. Accessed: 2015-11-02.

[138] "R lme4 archive." https://mirrors.ebi.ac.uk/CRAN/src/contrib/Archive/lme4/. Accessed: 2015-11-02.

[139] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[140] N. Horton and S. Lipsitz, "Multiple imputation in practice," *The American Statistician*, vol. 55, no. 3, pp. 244–254, 2001.

[141] "R mice archive." https://mirrors.ebi.ac.uk/CRAN/src/contrib/Archive/mice/. Accessed: 2015-11-02.

[142] P. S. 18, "Release version 18.0.0 (spss, inc., 2009, chicago, il, www.spss.com),"

[143] P. Royston, "Multiple imputation of missing values," *Stata Journal*, vol. 4, pp. 227–241, September 2004.

[144] S. Press, *Multiple-Imputation Reference Manual*. Stata Press, 2011.

[145] V. Kristman, M. Manno, and P. Cote, "Methods to account for attrition in longitudinal data: do they work? A simulation study," *European journal of epidemiology*, vol. 20, no. 8, pp. 657–662, 2005.

[146] M. Ali and O. Siddiqui, "Multiple imputation compared with some informative dropout procedures in the estimation and comparison of rates of change in longitudinal clinical trials with dropouts.," *Journal of biopharmaceutical statistics*, vol. 10, no. 2, p. 165, 2000.

[147] K. F. Schulz, D. G. Altman, D. Moher, *et al.*, "Consort 2010 statement: updated guidelines for reporting parallel group randomised trials," *BMC medicine*, vol. 8, no. 1, p. 18, 2010.

[148] I. Guideline, "Statistical principles for clinical trials," vol. 9, pp. 1905–1942, 1999.

[149] M. Liu, L. Wei, and J. Zhang, "Review of guidelines and literature for handling missing data in longitudinal clinical trials with a case study," *Pharmaceutical Statistics*, vol. 5, no. 1, pp. 7–18, 2006.

[150] N. R. Council, *The Prevention and Treatment of Missing Data in Clinical Trials*. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press, 2010.

[151] R. J. Little, R. D'Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, *et al.*, "The prevention and treatment of missing data in clinical trials," *New England Journal of Medicine*, vol. 367, no. 14, pp. 1355–1360, 2012.

[152] E. Von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, J. P. Vandenbroucke, S. Initiative, *et al.*, "The strengthening the reporting of ob-

servational studies in epidemiology (strobe) statement: guidelines for reporting observational studies," *Preventive medicine*, vol. 45, no. 4, pp. 247–251, 2007.

[153] A. Wood, I. White, and S. Thompson, "Are missing outcome data adequately handled? a review of published randomized controlled trials in major medical journals," *Clinical Trials*, vol. 1, no. 4, p. 368, 2004.

[154] J. Peugh and C. Enders, "Missing data in educational research: A review of reporting practices and suggestions for improvement," *Review of educational research*, vol. 74, no. 4, p. 525, 2004.

[155] H. Jeličić, E. Phelps, and R. Lerner, "Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology.," *Developmental Psychology*, vol. 45, no. 4, p. 1195, 2009.

[156] A. Mackinnon, "The use and reporting of multiple imputation in medical research–a review," *Journal of internal medicine*, 2010.

[157] A. Douiri, A. G. Rudd, and C. D. Wolfe, "Prevalence of poststroke cognitive impairment south london stroke register 1995–2010," *Stroke*, vol. 44, no. 1, pp. 138–145, 2013.

[158] C. Wolfe, S. Crichton, P. Heuschmann, C. McKevitt, A. Toschke, A. Grieve, and A. Rudd, "Estimates of outcomes up to ten years after stroke: Analysis from the prospective south london stroke register," *PLoS medicine*, vol. 8, no. 5, p. e1001033, 2011.

[159] M. Patel, C. Coshall, A. G. Rudd, and C. D. Wolfe, "Natural history of cognitive impairment after stroke and factors associated with its recovery," *Clinical Rehabilitation*, vol. 17, no. 2, pp. 158–166, 2003.

[160] D. Harari, C. Coshall, A. G. Rudd, and C. D. Wolfe, "New-onset fecal incontinence after stroke prevalence, natural history, risk factors, and impact," *Stroke*, vol. 34, no. 1, pp. 144–150, 2003.

[161] M. Patel, C. Coshall, E. Lawrence, A. G Rudd, and C. D. Wolfe, "Recovery from poststroke urinary incontinence: associated factors and impact on outcome," *Journal of the American Geriatrics Society*, vol. 49, no. 9, pp. 1229–1233, 2001.

[162] L. Ayerbe, S. Ayis, S. Crichton, C. D. Wolfe, and A. G. Rudd, "The natural history of depression up to 15 years after stroke the south london stroke register," *Stroke*, 2013.

[163] A. Sheldenkar, S. Crichton, A. Douiri, A. G. Rudd, C. D. Wolfe, and R. Chen, "Temporal trends in health-related quality of life after stroke: analysis from the south london stroke register 1995–2011," *International Journal of Stroke*, vol. 9, no. 6, pp. 721–727, 2014.

[164] C. McKevitt, R. Dundas, C. Wolfe, *et al.*, "Two simple questions to assess outcome after stroke a european study," *Stroke*, vol. 32, no. 3, pp. 681–686, 2001.

[165] N. S. Graham, S. Crichton, M. Koutroumanidis, C. D. Wolfe, and A. G. Rudd, "Incidence and associations of poststroke epilepsy the prospective south london stroke register," *Stroke*, vol. 44, no. 3, pp. 605–611, 2013.

[166] J. Redfern, C. McKevitt, R. Dundas, A. G. Rudd, and C. D. Wolfe, "Behavioral risk factor prevalence and lifestyle change after stroke a prospective study," *Stroke*, vol. 31, no. 8, pp. 1877–1881, 2000.

[167] T. Hillen, R. Dundas, E. Lawrence, J. A. Stewart, A. G. Rudd, and C. D. Wolfe, "Antithrombotic and antihypertensive management 3 months after ischemic stroke a prospective study in an inner city population," *Stroke*, vol. 31, no. 2, pp. 469–475, 2000.

[168] K. Aho, P. Harmsen, S. Hatano, J. Marquardsen, V. Smirnov, and T. Strasser, "Cerebrovascular disease in the community: results of a who collaborative study," *Bulletin of the World Health Organization*, vol. 58, no. 1, p. 113, 1980.

[169] A. G. Thrift, H. M. Dewey, R. A. Macdonell, J. J. McNeil, and G. A. Donnan, "Incidence of the major stroke subtypes initial findings from the north east melbourne stroke incidence study (nemesis)," *Stroke*, vol. 32, no. 8, pp. 1732–1738, 2001.

[170] I. S. W. Party, "National clinical guideline for stroke," 2008.

[171] H. S. Jørgensen, H. Nakayama, H. O. Raaschou, and T. S. Olsen, "Intracerebral hemorrhage versus infarction: stroke severity, risk factors, and prognosis," *Annals of neurology*, vol. 38, no. 1, pp. 45–50, 1995.

[172] Y. Wang, A. G. Rudd, and C. D. Wolfe, "Trends and survival between ethnic groups after stroke the south london stroke register," *Stroke*, vol. 44, no. 2, pp. 380–387, 2013.

[173] V. L. Feigin, M. H. Forouzanfar, R. Krishnamurthi, G. A. Mensah, M. Connor, D. A. Bennett, A. E. Moran, R. L. Sacco, L. Anderson, T. Truelsen, *et al.*, "Global and regional burden of stroke during 1990–2010: findings from the global burden of disease study 2010," *The Lancet*, vol. 383, no. 9913, pp. 245–255, 2014.

[174] ONS, "Deaths registered in England and Wales in 2010 by cause, Statistical Bulletin.," *Statistical Bulletin*, 2011.

[175] Ö. Saka, A. McGuire, and C. Wolfe, "Cost of stroke in the united kingdom," *Age and ageing*, vol. 38, no. 1, pp. 27–32, 2009.

[176] C. Sudlow and C. Warlow, "Comparing stroke incidence worldwide what makes studies comparable?," *Stroke*, vol. 27, no. 3, pp. 550–558, 1996.

[177] Y. Wang, A. G. Rudd, and C. D. Wolfe, "Age and ethnic disparities in incidence of stroke over time the south london stroke register," *Stroke*, vol. 44, no. 12, pp. 3298–3304, 2013.

[178] J. Stewart, R. Dundas, R. Howard, A. Rudd, and C. Wolfe, "Ethnic differences in incidence of stroke: prospective study with stroke register," *Bmj*, vol. 318, no. 7189, pp. 967–971, 1999.

[179] C. Wolfe, N. Smeeton, C. Coshall, K. Tilling, and A. Rudd, "Survival differences after stroke in a multiethnic population: follow-up study with the south london stroke register," *Bmj*, vol. 331, no. 7514, p. 431, 2005.

[180] C. Hajat, R. Dundas, J. A. Stewart, E. Lawrence, A. G. Rudd, R. Howard, and C. D. Wolfe, "Cerebrovascular risk factors and stroke subtypes differences between ethnic groups," *Stroke*, vol. 32, no. 1, pp. 37–42, 2001.

[181] J. Addo, A. Bhalla, S. Crichton, A. Rudd, C. McKevitt, and C. Wolfe, "Provision of acute stroke care and associated factors in a multiethnic population: prospective study with the south london stroke register," *BMJ: British Medical Journal*, vol. 342, 2011.

[182] K. M. Mohan, S. L. Crichton, A. P. Grieve, A. G. Rudd, C. D. A. Wolfe, and P. U. Heuschmann, "Frequency and predictors for the risk of stroke recurrence up to 10 years after stroke: the south london stroke register," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 80, no. 9, pp. 1012–1018, 2009.

[183] "Report of a who expert committee. world health organ tech report," *Arterial hypertension*, 1978.

[184] P. Heuschmann, A. Grieve, A. Toschke, A. Rudd, and C. Wolfe, "Ethnic group disparities in 10-year trends in stroke incidence and vascular risk factors," *Stroke*, vol. 39, no. 8, pp. 2204–2210, 2008.

[185] D. Wade and C. Collin, "The Barthel ADL Index: a standard measure of physical disability?," *Disability & Rehabilitation*, vol. 10, no. 2, pp. 64–67, 1988.

[186] C. Wolfe, N. Taub, E. Woodrow, and P. Burney, "Assessment of scales of disability and handicap for stroke patients," *Stroke*, vol. 22, no. 10, p. 1242, 1991.

[187] D. Wade, J. Legh-Smith, and R. Hewer, "Social activities after stroke: measurement and natural history using the Frenchay Activities Index," *Disability & Rehabilitation*, vol. 7, no. 4, pp. 176–181, 1985.

[188] C. S. Anderson, K. D. Jamrozik, R. J. Broadhurst, and E. G. Stewart-Wynne, "Predicting survival for 1 year among different subtypes of stroke. results from the perth community stroke study.," *Stroke*, vol. 25, no. 10, pp. 1935–1944, 1994.

[189] A. Zigmond and R. Snaith, "The hospital anxiety and depression scale," *Acta Psychiatrica Scandinavica*, vol. 67, no. 6, pp. 361–370, 1983.

[190] I. Aben, F. Verhey, R. Lousberg, J. Lodder, and A. Honig, "Validity of the beck depression inventory, hospital anxiety and depression scale, scl-90, and hamilton depression rating scale as screening instruments for depression in stroke patients," *Psychosomatics*, vol. 43, no. 5, pp. 386–393, 2002.

[191] I. Bjelland, A. Dahl, T. Haug, and D. Neckelmann, "The validity of the Hospital Anxiety and Depression Scale:: An updated literature review," *Journal of psychosomatic Research*, vol. 52, no. 2, pp. 69–77, 2002.

[192] M. Folstein, S. Folstein, and P. McHugh, "'mini-mental state': A practical method of grading the cognitive state of patients for the clinician.," *Journal of Psychiatric Research*, vol. 12, pp. 189–98, 1975.

[193] H. Hodkinson, "Evaluation of a mental test score for assessment of mental impairment in the elderly," *Age and ageing*, vol. 1, no. 4, p. 233, 1972.

[194] J. P. Tuijl, E. M. Scholte, A. J. Craen, and R. C. Mast, "Screening for cognitive impairment in older general hospital patients: comparison of the six-item cog-

nitive impairment test with the mini-mental state examination," *International journal of geriatric psychiatry*, vol. 27, no. 7, pp. 755–762, 2012.

[195] S. Jitapunkul, I. Pillay, and S. Ebrahim, "The abbreviated mental test: its use and validity," *Age and ageing*, vol. 20, no. 5, p. 332, 1991.

[196] T. Tombaugh and N. McIntyre, "The mini-mental state examination: a comprehensive review.," *Journal of the American Geriatrics Society*, vol. 40, no. 9, p. 922, 1992.

[197] J. E. Ware, M. Kosinski, and S. Keller, *SF-36 physical and mental health summary scales: a user's manual*. Health Assessment Lab., 1994.

[198] J. E. Ware, M. Kosinski, and S. D. Keller, *SF-12: How to score the SF-12 physical and mental health summary scales*. Health Institute, New England Medical Center, 1995.

[199] A. S. Pickard, J. A. Johnson, A. Penn, F. Lau, and T. Noseworthy, "Replicability of sf-36 summary scores by the sf-12 in stroke patients," *Stroke*, vol. 30, no. 6, pp. 1213–1217, 1999.

[200] K. Tilling, J. A. Sterne, A. G. Rudd, T. A. Glass, R. J. Wityk, and C. D. Wolfe, "A new method for predicting recovery after stroke," *Stroke*, vol. 32, no. 12, pp. 2867–2873, 2001.

[201] A. Toschke, K. Tilling, A. Cox, A. Rudd, P. Heuschmann, and C. Wolfe, "Patient-specific recovery patterns over time measured by dependence in activities of daily living after stroke and post-stroke care: The south london stroke register (slsr)," *European Journal of Neurology*, vol. 17, no. 2, pp. 219–225, 2010.

[202] L. Ayerbe, S. A. Ayis, S. Crichton, C. D. Wolfe, and A. G. Rudd, "Natural history, predictors and associated outcomes of anxiety up to 10 years after stroke: the south london stroke register," *Age and ageing*, p. aft208, 2013.

[203] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[204] S. Lacus, *RRP*, 2009. R package version 2.9.0.

[205] J. Pinheiro and D. Bates, *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media, 2006.

[206] "Sas 9.3 for windows, sas institute, cary (2011),"

[207] A. Touloumis, "R package multgee: A generalized estimating equations solver for multinomial responses," *Journal of Statistical Software*, 2015.

[208] L. C. Thygesen, C. Johansen, N. Keiding, E. Giovannucci, and M. Grønbæk, "Effects of sample attrition in a longitudinal study of the association between alcohol intake and all-cause mortality," *Addiction*, vol. 103, no. 7, pp. 1149–1159, 2008.

[209] P. Lavikainen, E. Leskinen, S. Hartikainen, J. Möttönen, R. Sulkava, and M. J. Korhonen, "Impact of missing data mechanism on the estimate of change: a case study on cognitive function and polypharmacy among older persons," *Clinical epidemiology*, vol. 7, p. 169, 2015.

[210] K. B. Rajan and S. E. Leurgans, "Joint modeling of missing data due to non-participation and death in longitudinal aging studies," *Statistics in medicine*, vol. 29, no. 21, pp. 2260–2268, 2010.

[211] M. Kauppi, T. Sokka, and P. Hannonen, "Survey nonresponse is associated with increased mortality in patients with rheumatoid arthritis and in a community population.," *The Journal of rheumatology*, vol. 32, no. 5, pp. 807–810, 2005.

[212] R. Curtin, S. Presser, and E. Singer, "Changes in telephone survey nonresponse over the past quarter century," *Public opinion quarterly*, vol. 69, no. 1, pp. 87–98, 2005.

[213] C. L. Booker, S. Harding, and M. Benzeval, "A systematic review of the effect of retention methods in population-based cohort studies," *BMC Public Health*, vol. 11, no. 1, p. 249, 2011.

[214] F. Barzi, M. Woodward, R. M. Marfisi, G. Tognoni, R. Marchioli, G.-P. Investigators, *et al.*, "Analysis of the benefits of a mediterranean diet in the gissi-prevenzione study: a case study in imputation of missing values from repeated measurements," *European journal of epidemiology*, vol. 21, no. 1, pp. 15–24, 2006.

[215] L. D. Howe, K. Tilling, B. Galobardes, and D. A. Lawlor, "Loss to follow-up in cohort studies: bias in estimates of socioeconomic inequalities," *Epidemiology*, vol. 24, no. 1, pp. 1–9, 2013.

[216] G. Encrenaz, V. Rondeau, A. Messiah, and M. Auriacombe, "Examining the influence of drop-outs in a follow-up of maintained opiate users," *Drug and alcohol dependence*, vol. 79, no. 3, pp. 303–310, 2005.

[217] S. A. Peters, M. L. Bots, H. M. den Ruijter, M. K. Palmer, D. E. Grobbee, J. R. Crouse, D. H. OLeary, G. W. Evans, J. S. Raichlen, K. G. Moons, *et al.*, "Multiple imputation of missing repeated outcome measurements did not add to linear mixed-effects models," *Journal of clinical epidemiology*, vol. 65, no. 6, pp. 686–695, 2012.

[218] J. Twisk, M. de Boer, W. de Vente, and M. Heymans, "Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis," *Journal of clinical epidemiology*, vol. 66, no. 9, pp. 1022–1028, 2013.

[219] A. Burton, D. Altman, P. Royston, and R. Holder, "The design of simulation studies in medical statistics," *Statistics in Medicine*, vol. 25, no. 24, pp. 4279–4292, 2006.

[220] I. R. White and J. B. Carlin, "Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values," *Statistics in medicine*, vol. 29, no. 28, pp. 2920–2931, 2010.

[221] Å. M. Johansson and M. O. Karlsson, "Comparison of methods for handling missing covariate data," *The AAPS journal*, vol. 15, no. 4, pp. 1232–1241, 2013.

[222] B. F. Kurland, L. L. Johnson, B. L. Egleston, and P. H. Diehr, "Longitudinal data with follow-up truncated by death: match the analysis method to research aims," *Statistical science: a review journal of the Institute of Mathematical Statistics*, vol. 24, no. 2, p. 211, 2009.

[223] O. Harel, S. M. Hofer, L. Hoffman, N. L. Pedersen, and B. Johansson, "Population inference with mortality and attrition in longitudinal studies on aging: A two-stage multiple imputation method," *Experimental aging research*, vol. 33, no. 2, pp. 187–203, 2007.

[224] O. Harel and H. Demirtas, "Re: Joint modeling of missing data due to non-participation and death in longitudinal aging studies," *Statistics in medicine*, vol. 30, no. 21, pp. 2663–2665, 2011.

[225] S. L. Brilleman, N. A. Pachana, and A. J. Dobson, "The impact of attrition on the representativeness of cohort studies of older people," *BMC medical research methodology*, vol. 10, no. 1, p. 71, 2010.

[226] G. Fraser and R. Yan, "Guided multiple imputation of missing data: using a subsample to strengthen the missing-at-random assumption," *Epidemiology*, vol. 18, no. 2, pp. 246–252, 2007.

[227] H. Demirtas, "Multiple imputation under bayesianly smoothed pattern-mixture models for non-ignorable drop-out," *Statistics in Medicine*, vol. 24, no. 15, pp. 2345–2363, 2005.

[228] H. Demirtas, "Simulation driven inferences for multiply imputed longitudinal datasets*," *Statistica Neerlandica*, vol. 58, no. 4, pp. 466–482, 2004.

[229] G. Molenberghs, H. Thijs, M. G. Kenward, and G. Verbeke, "Sensitivity analysis of continuous incomplete longitudinal outcomes," *Statistica Neerlandica*, vol. 57, no. 1, pp. 112–135, 2003.

[230] S. Mazumdar, G. Tang, P. R. Houck, M. A. Dew, A. E. Begley, J. Scott, B. H. Mulsant, and C. F. Reynolds, "Statistical analysis of longitudinal psychiatric data with dropouts," *Journal of psychiatric research*, vol. 41, no. 12, pp. 1032–1041, 2007.

[231] D. A. Newman, "Missing data five practical guidelines," *Organizational Research Methods*, vol. 17, no. 4, pp. 372–411, 2014.

[232] G. R. Locke III, C. D. Schleck, J. Y. Ziegenfuss, T. J. Beebe, A. R. Zinsmeister, N. J. Talley, *et al.*, "A low response rate does not necessarily indicate non-response bias in gastroenterology survey research: a population-based study," *Journal of Public Health*, vol. 21, no. 1, pp. 87–95, 2013.

[233] R. J. Lacey, K. P. Jordan, and P. R. Croft, "Does attrition during follow-up of a population cohort study inevitably lead to biased estimates of health status?," *PloS one*, vol. 8, no. 12, p. e83948, 2013.

Appendix A

R and SAS code

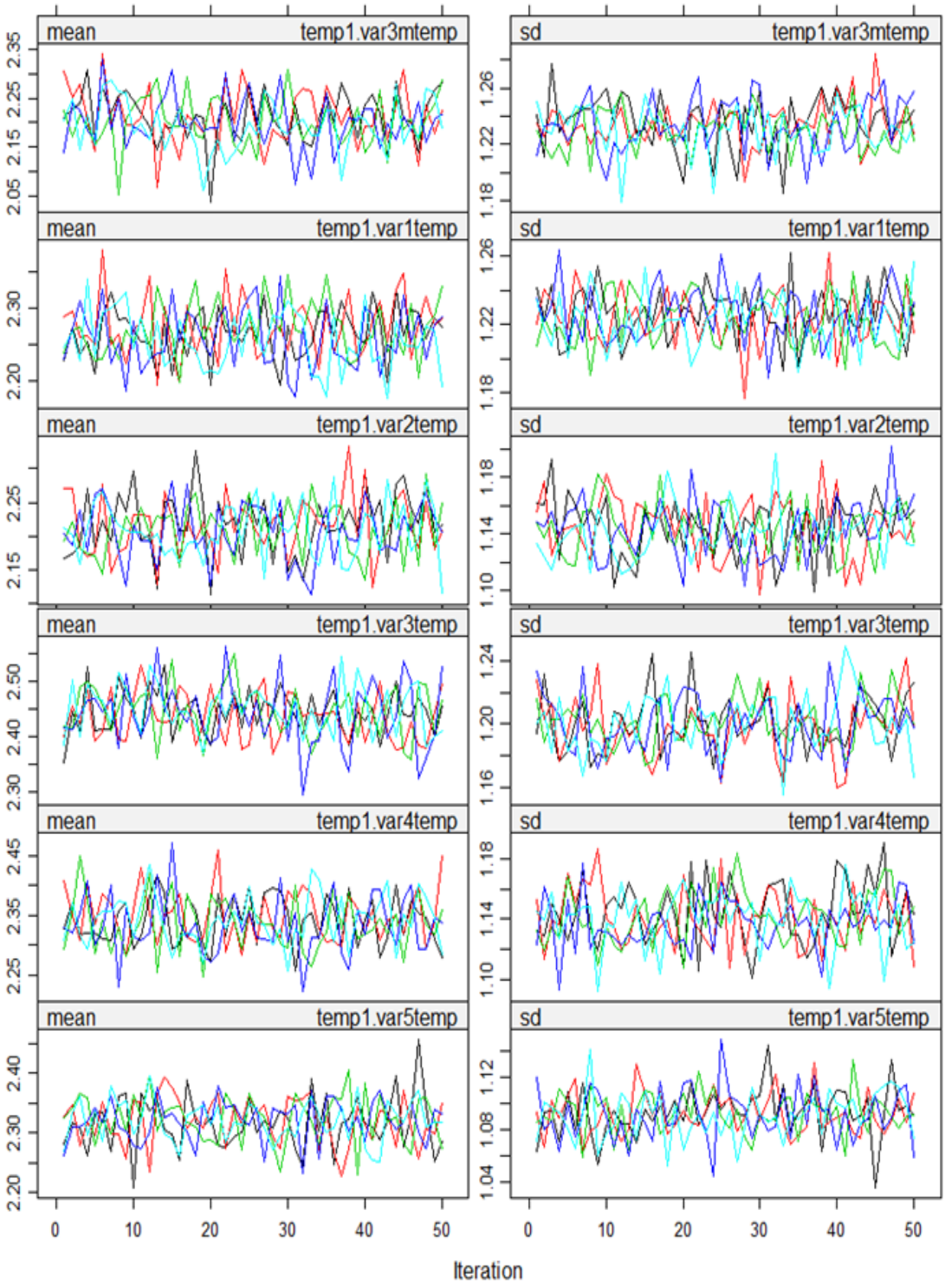
Copies of the R and SAS code used in the analyses presented in Chapters 6 and 7 are provided on the CD attached to the back cover of the thesis.

R code used to run the four scenarios explored in the simulation study (Chapter 6) is included in the folder titled "Simulation study" with a separate file for each scenario.

Codes for the models applied in Chapter 7 are stored in the folder titled "Longitudinal models". Within this folder there are individual files for each model type (i.e GEE, mortal WGEE, immortal WGEE, MIGEE, GLMM, shared parameter and pattern mixture models). All models were fitted using SAS with the exception of the multinomial GEE which was fitted using R.

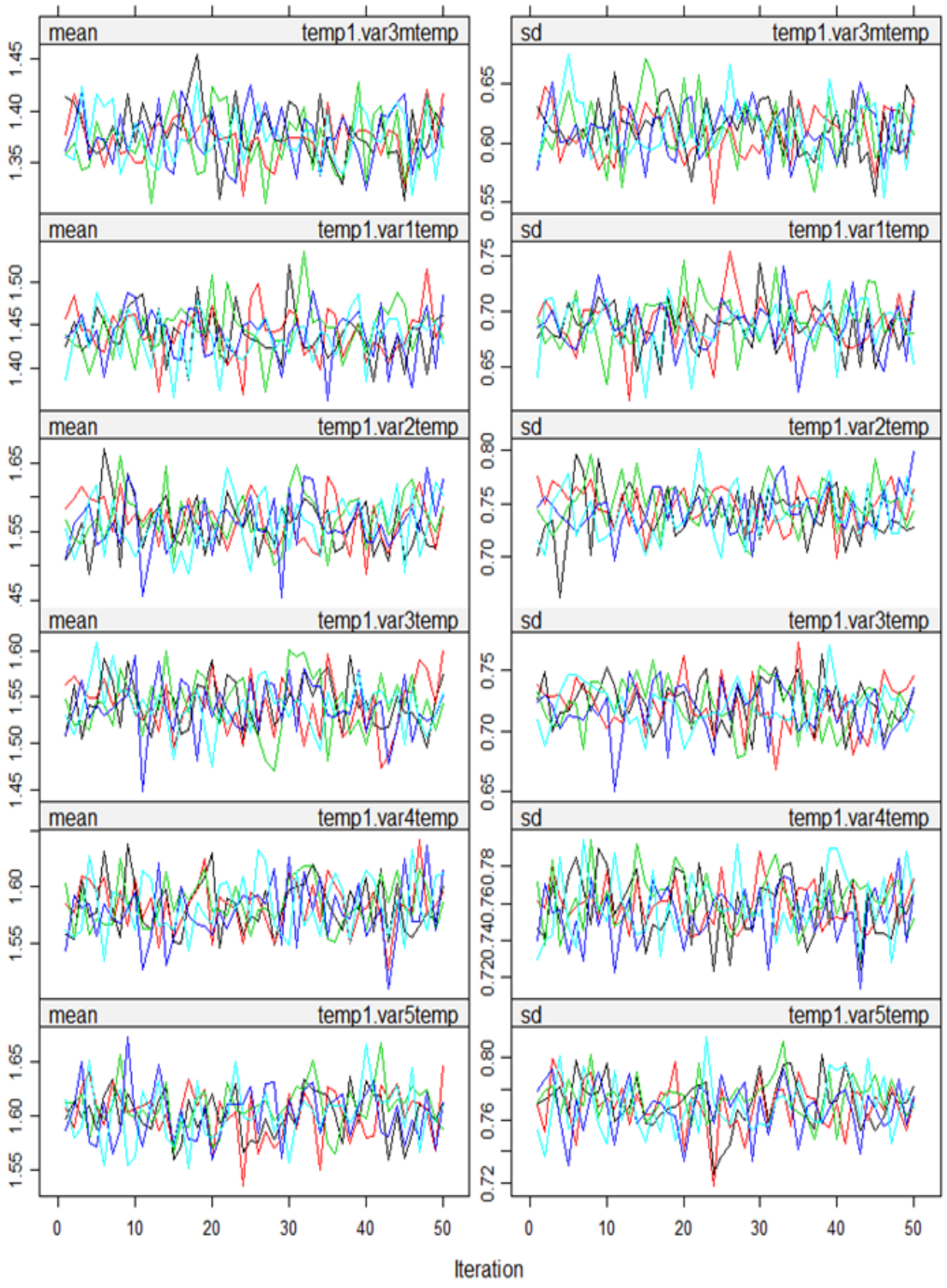
Appendix B

Imputed values by iteration number using MICE



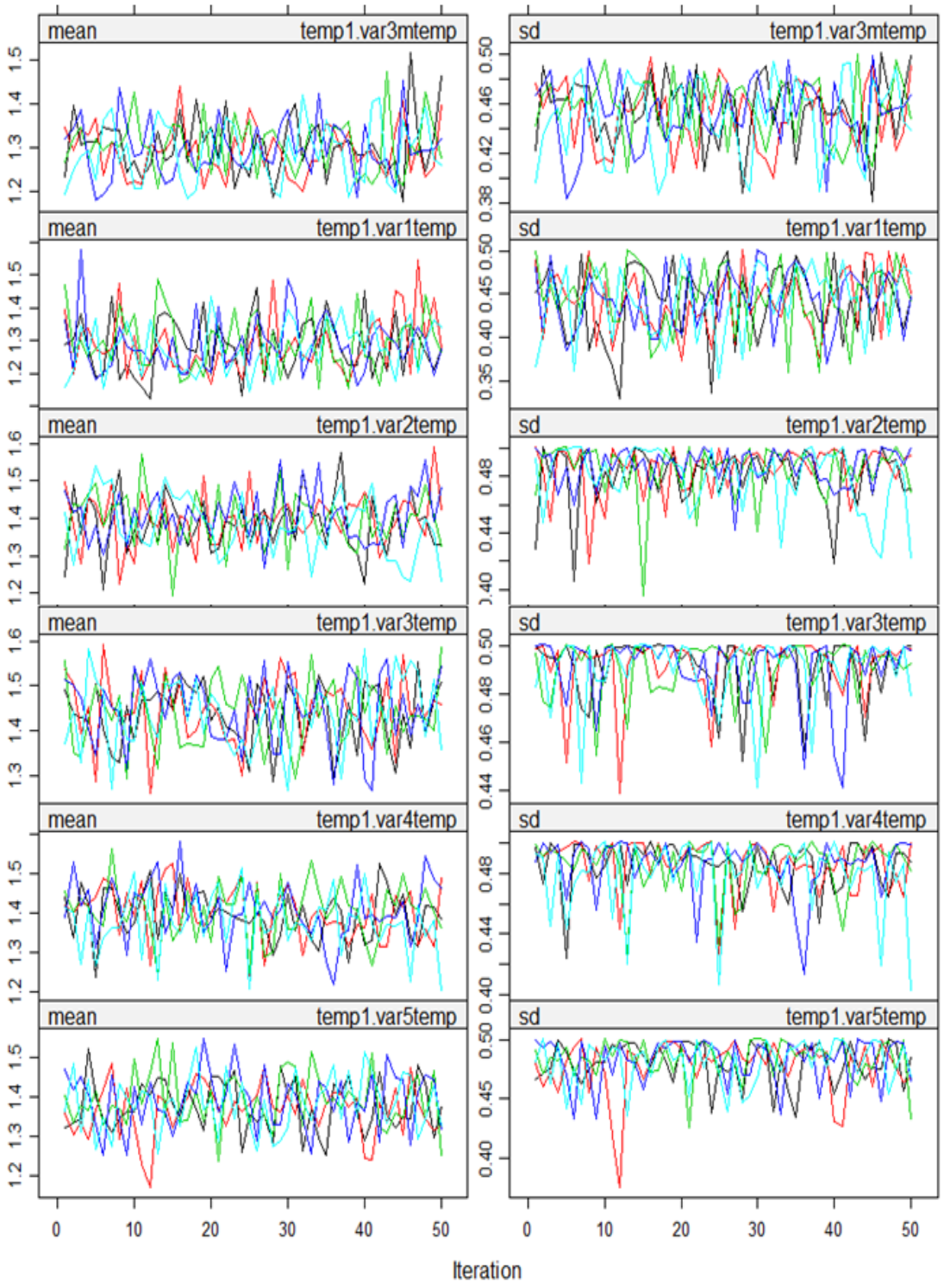
Note: Mean and standard deviation of imputed values across using MICE with five chains and 50 iterations per chain

Figure B.1: Mean and standard deviation of imputed values of categorical disability level variables



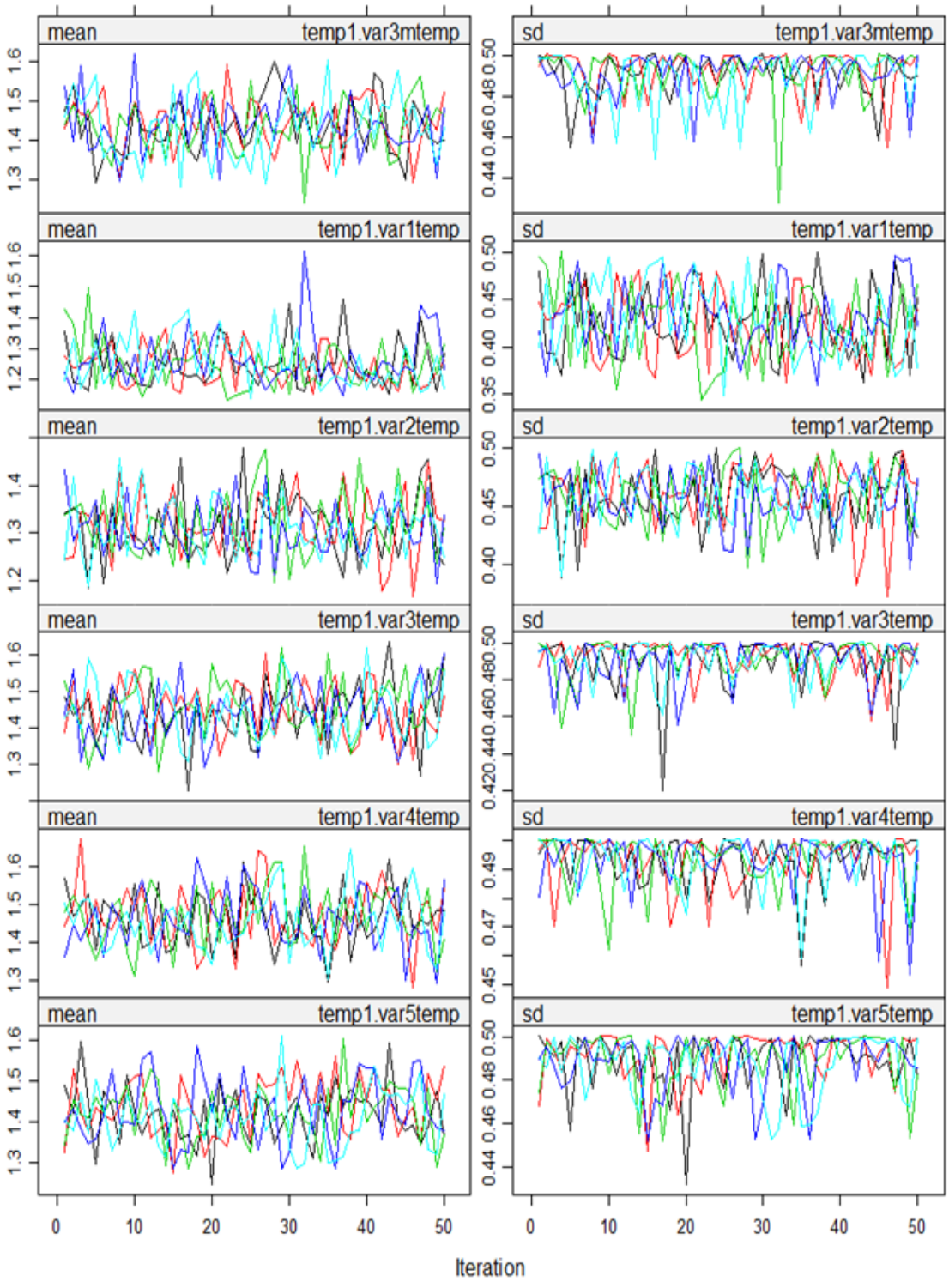
Note: Mean and standard deviation of imputed values across using MICE with five chains and 50 iterations per chain

Figure B.2: Mean and standard deviation of imputed values of categorical activity level variables



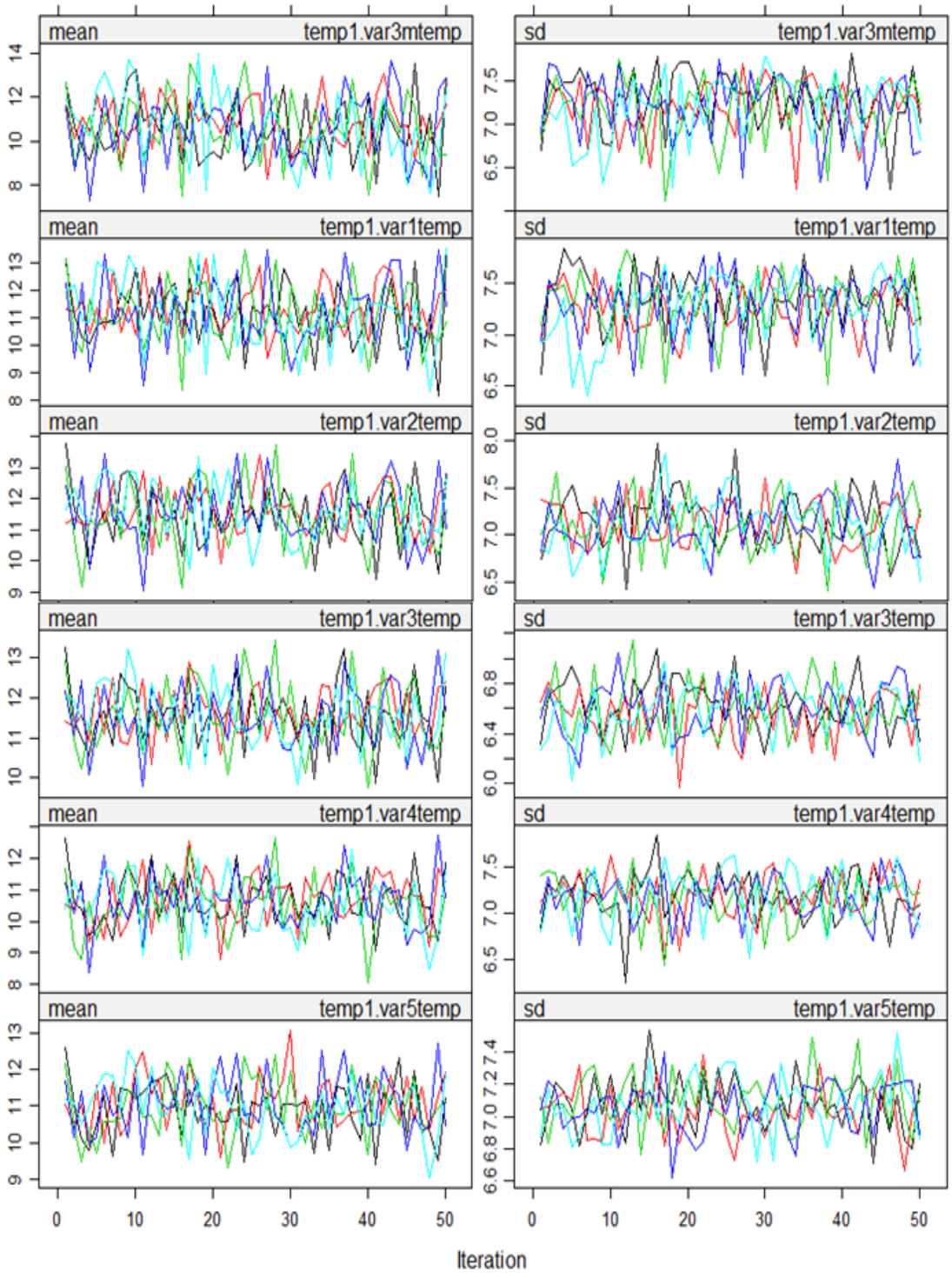
Note: Mean and standard deviation of imputed values across using MICE with five chains and 50 iterations per chain

Figure B.3: Mean and standard deviation of imputed values of binary anxiety variables



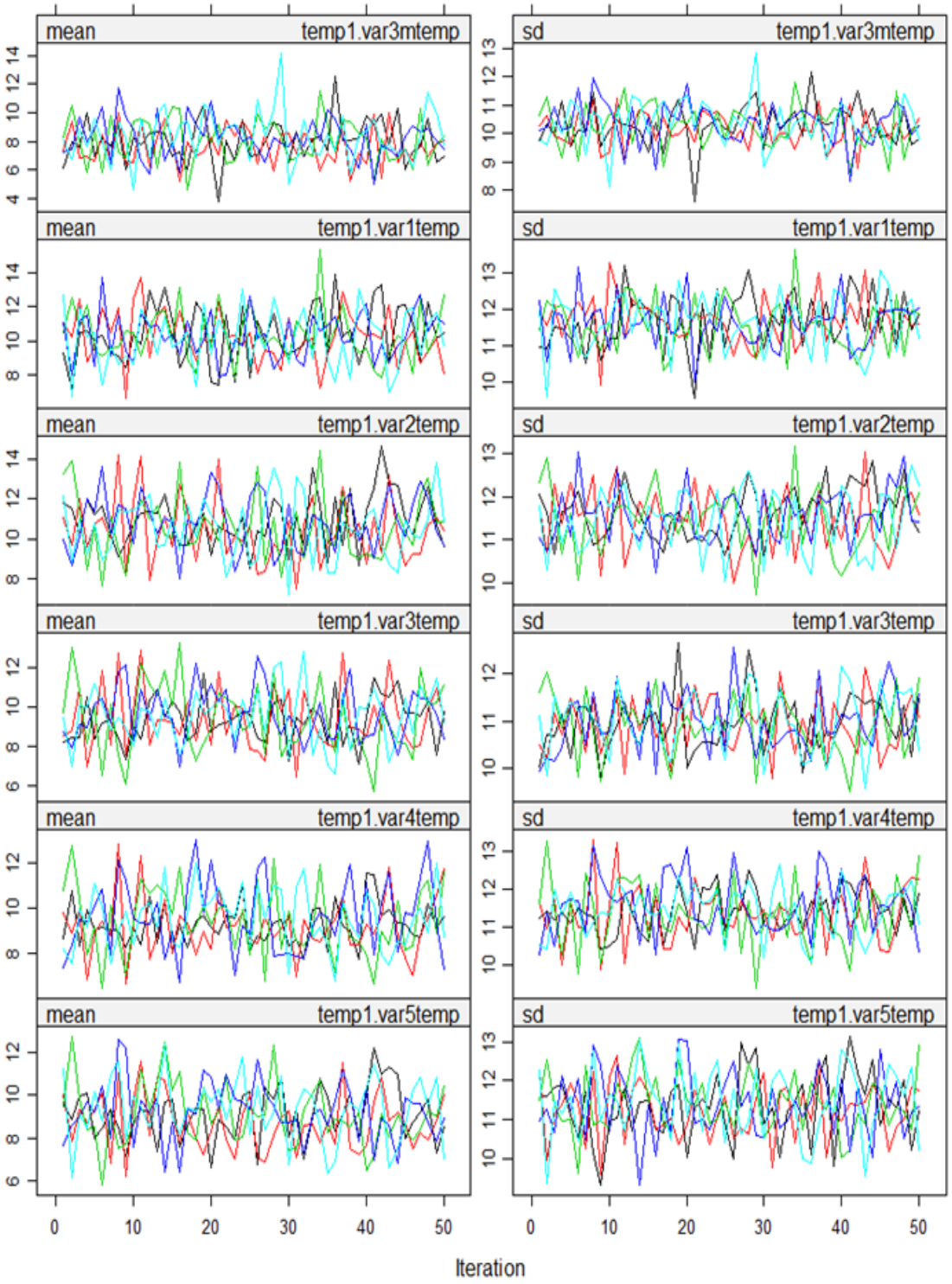
Note: Mean and standard deviation of imputed values across using MICE with five chains and 50 iterations per chain

Figure B.4: Mean and standard deviation of imputed values of binary depression variables



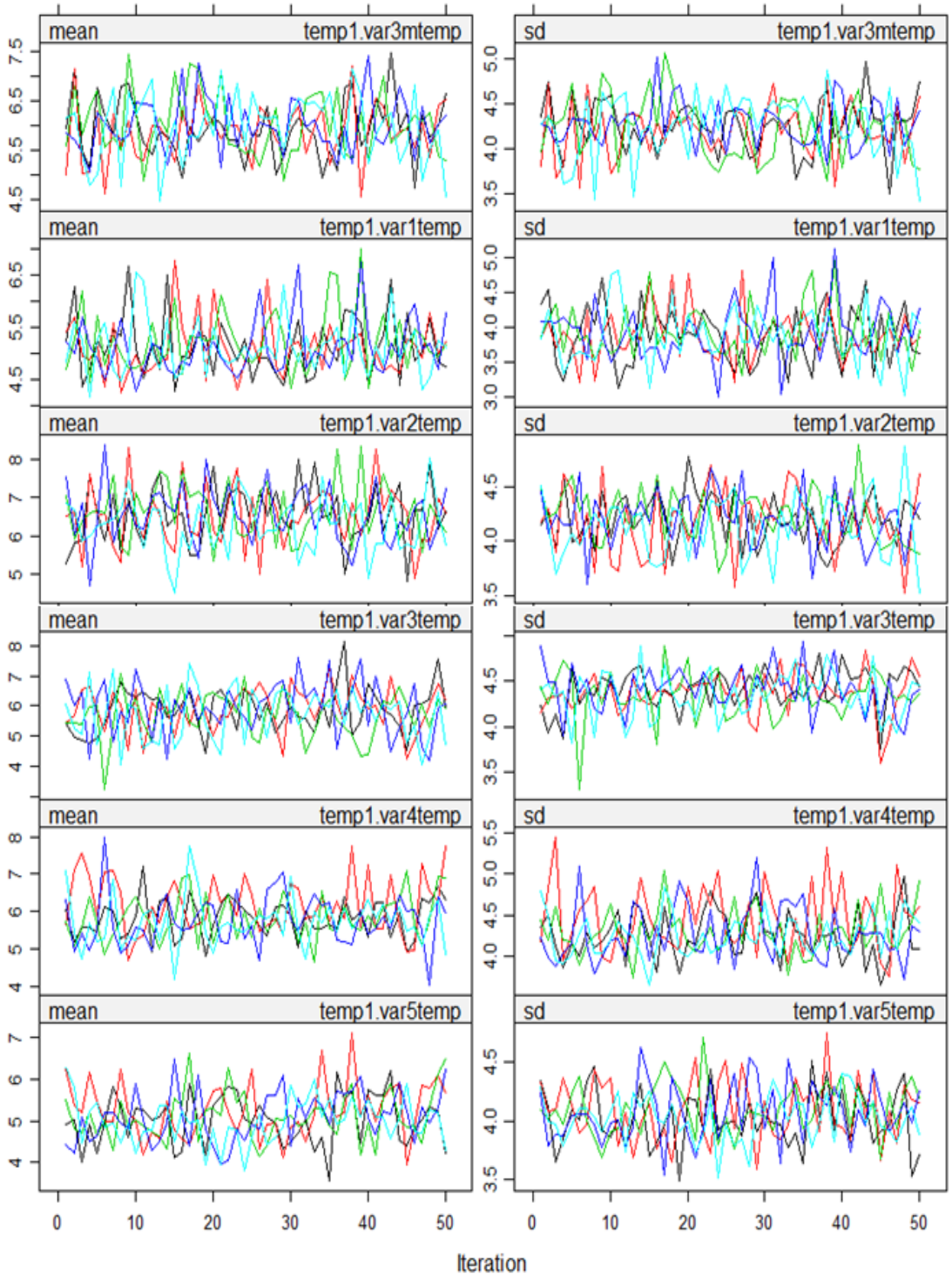
Note: Mean and standard deviation of imputed values across using MICE with five chains and 50 iterations per chain

Figure B.5: Mean and standard deviation of imputed values of Barthel Index score



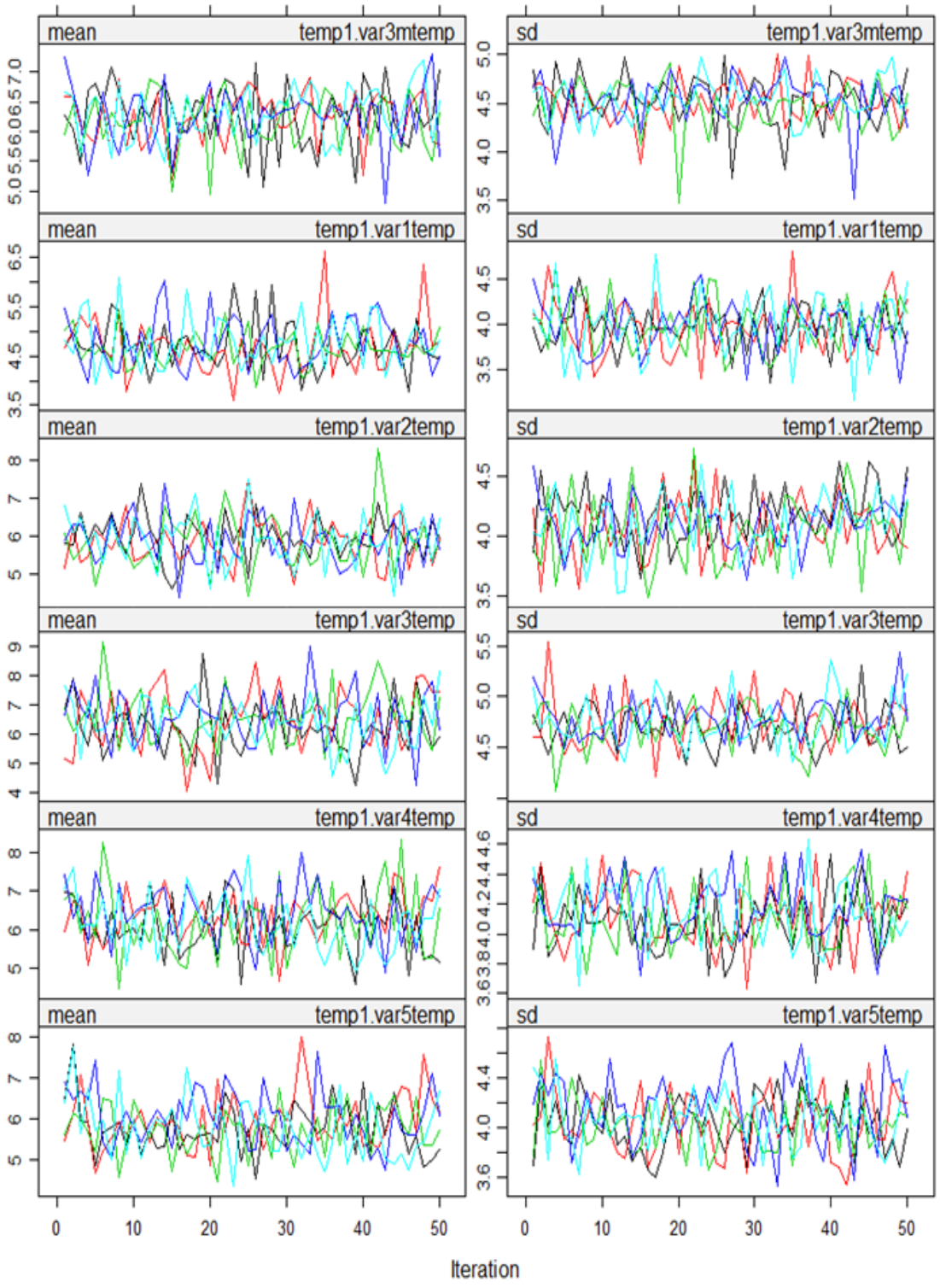
Note: Mean and standard deviation of imputed values across using MICE with five chains and 50 iterations per chain

Figure B.6: Mean and standard deviation of imputed values of Frenchay Activities Index score



Note: Mean and standard deviation of imputed values across using MICE with five chains and 50 iterations per chain

Figure B.7: Mean and standard deviation of imputed values of HADs - anxiety score

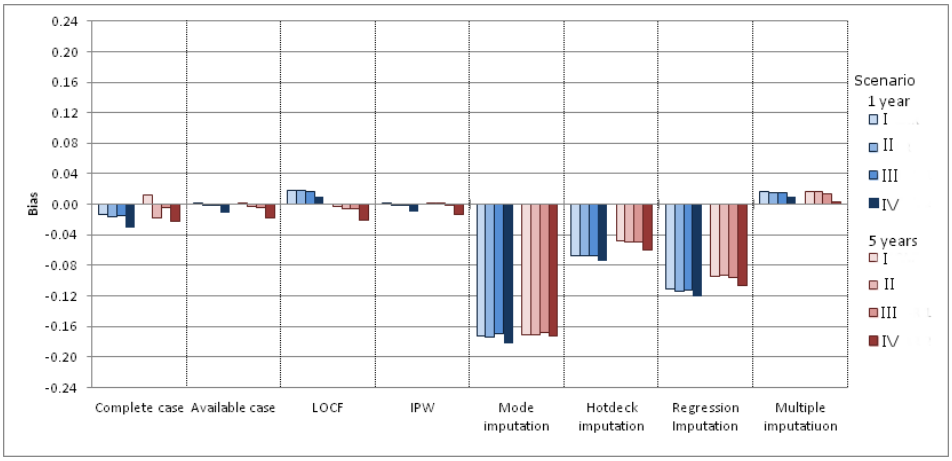


Note: Mean and standard deviation of imputed values across using MICE with five chains and 50 iterations per chain

Figure B.8: Mean and standard deviation of imputed values of HADs - depression score

Appendix C

Impact of missing data on estimates of prevalence of anxiety

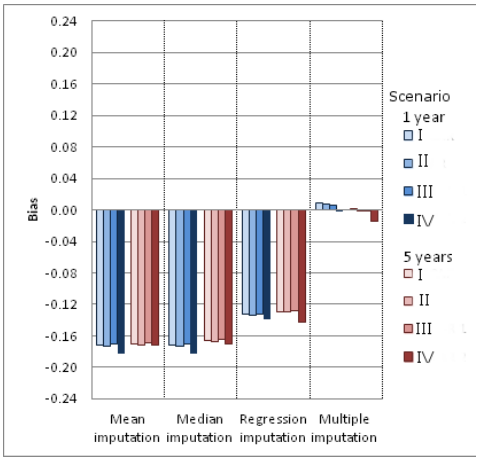


Note: Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses with the following exceptions: complete case analysis included mean n=277 at one year and n=100 at five years; available case and IPW included mean=337 and n=206 and LOCF used data from mean n=419 and n=302 with data recorded at at least one previous follow-up.

Mean prevalence rate in complete data=31% at one year and 32% at five years.

Abbreviations: LOCF last observation carried forward, IPW inverse probability weighting

Figure C.1: Bias of estimates of prevalence of anxiety using categorical missing data methods



Note: Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses. Mean prevalence rate in complete data=31% at one year and 32% at five years.

Figure C.2: Bias of estimates of prevalence of anxiety using continuous imputation methods

Table C.1: Bias associated with missing data methods when estimating prevalence of anxiety

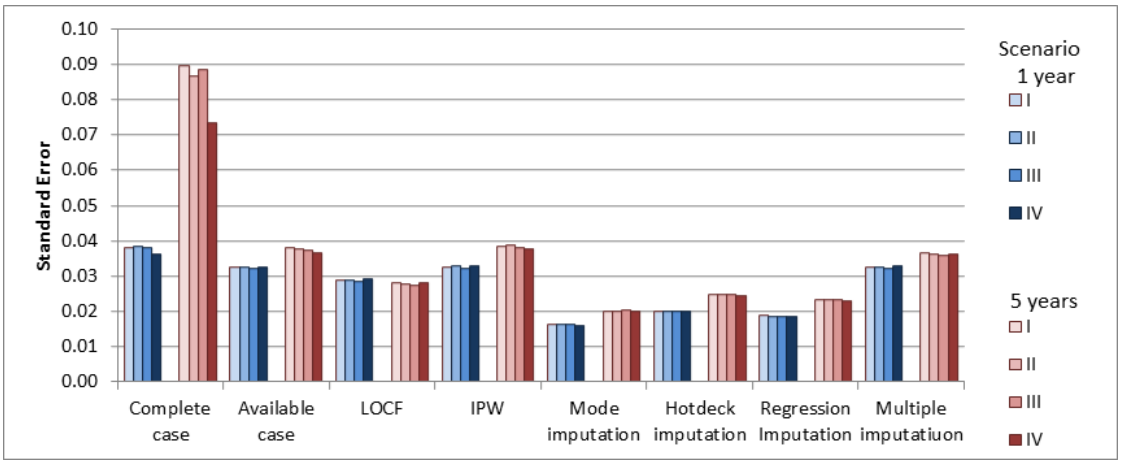
		1 year				5 years			
		Sc I	Sc II	Sc III	Sc IV	Sc I	Sc II	Sc III	Sc IV
Categorical	Complete case	-0.014	-0.016	-0.015	-0.031	0.012	-0.018	-0.004	-0.022
	Available case	0.001	-0.000	-0.001	-0.011	0.000	-0.004	-0.005	-0.018
	LOCF	0.017	0.018	0.017	0.010	-0.003	-0.006	-0.006	-0.021
	IPW	0.002	-0.000	-0.000	-0.009	0.001	0.002	-0.001	-0.013
	Mode imputation	-0.172	-0.174	-0.170	-0.182	-0.170	-0.172	-0.169	-0.172
	Hotdeck imputation	-0.067	-0.068	-0.068	-0.074	-0.048	-0.049	-0.050	-0.060
	Regression imputation	-0.111	-0.114	-0.113	-0.120	-0.095	-0.093	-0.095	-0.107
	Multiple imputatiuon	0.017	0.015	0.015	0.010	0.017	0.016	0.014	0.004
Continuous	Median imputation	-0.172	-0.174	-0.171	-0.182	-0.171	-0.172	-0.170	-0.172
	Median imputation	-0.172	-0.174	-0.170	-0.182	-0.166	-0.167	-0.165	-0.171
	Hotdeck imputation	-0.049	-0.049	-0.049	-0.057	-0.033	-0.033	-0.035	-0.049
	Regression imputation	-0.132	-0.134	-0.133	-0.139	-0.130	-0.130	-0.128	-0.142
	Multiple imputation	0.008	0.007	0.006	-0.001	0.000	-0.001	-0.002	-0.0134

Scenario I assumes outcome data are MCAR, Scenario II Assumes MAR, Scenario III assumed MNAR with missingness dependent on time of death and Scenario IV assumes MNAR data with missingness dependent on current disability level

Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses with the following exceptions: complete case analysis included mean n=277 at one year and n=100 at five years; available case and IPW included mean=337 and n=206 and LOCF used data from mean n=419 and n=302 with data recorded at at least one previous follow-up.

Mean prevalence rate in complete data=31% at one year and 32% at five years.

Abbreviations: LOCF last observation carried forward, IPW Inverse probability weighting, MCAR missing completely at random, MAR missing at random, MNAR missing not at random.

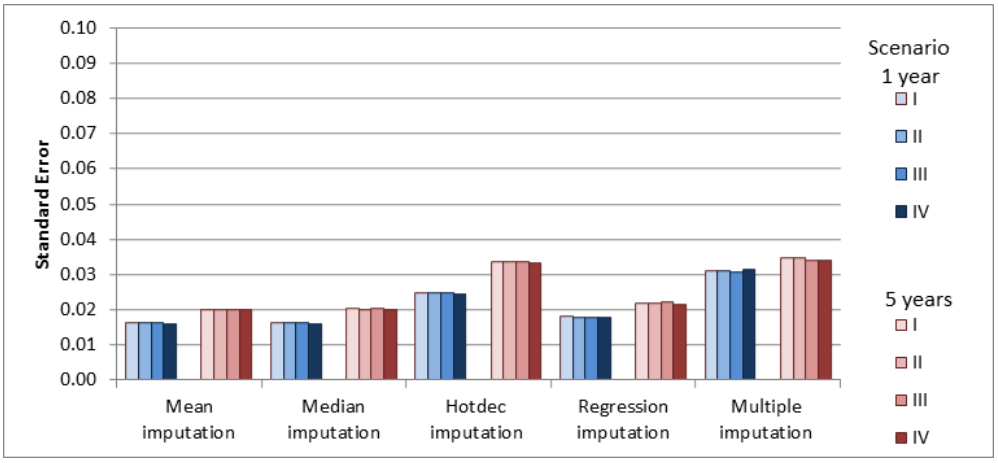


Note: Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses with the following exceptions: complete case analysis included mean n=277 at one year and n=100 at five years; available case and IPW included mean=337 and n=206 and LOCF used data from mean n=419 and n=302 with data recorded at at least one previous follow-up.

Mean prevalence rate in complete data=31% at one year and 32% at five years.

Abbreviations: LOCF last observation carried forward, IPW inverse probability weighting

Figure C.3: Standard error of estimates of prevalence of anxiety using categorical missing data methods



Note: Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses. Mean prevalence rate in complete data=31% at one year and 32% at five years.

Figure C.4: Standard error of estimates of prevalence of anxiety using continuous imputation methods

Table C.2: Precision of methods when estimating prevalence of anxiety

		1 year				5 years			
		Sc I	Sc II	Sc III	Sc IV	Sc I	Sc II	Sc III	Sc IV
Categorical	True se	0.028	0.028	0.028	0.028	0.029	0.029	0.029	0.029
	Complete case	0.038	0.038	0.038	0.036	0.090	0.087	0.089	0.073
	Available case	0.032	0.033	0.0321	0.0331	0.038	0.038	0.037	0.037
	LOCF	0.029	0.029	0.028	0.029	0.028	0.028	0.027	0.028
	IPW	0.033	0.033	0.032	0.033	0.039	0.039	0.038	0.038
	Mode imputation	0.016	0.016	0.016	0.016	0.020	0.020	0.020	0.020
	Hotdeck imputation	0.020	0.020	0.020	0.019	0.025	0.025	0.025	0.025
	Regression imputation	0.019	0.019	0.019	0.018	0.023	0.023	0.023	0.023
	Multiple imputatiuon	0.033	0.032	0.032	0.033	0.037	0.036	0.036	0.036
Continuous	Median imputation	0.016	0.016	0.016	0.016	0.020	0.020	0.0202	0.020
	Median imputation	0.016	0.016	0.016	0.016	0.020	0.020	0.020	0.020
	Hotdeck imputation	0.025	0.025	0.025	0.025	0.034	0.034	0.034	0.033
	Regression imputation	0.018	0.018	0.018	0.018	0.022	0.022	0.022	0.022
	Multiple imputation	0.031	0.031	0.031	0.032	0.035	0.035	0.034	0.034

Scenario I assumes outcome data are MCAR, Scenario II Assumes MAR, Scenario III assumed MNAR with missingness dependent on time of death and Scenario IV assumes MNAR data with missingness dependent on current disability level Mean number alive at follow-up was 450 at one year and 319 at five years and were included in analyses with the following exceptions: complete case analysis included mean n=277 at one year and n=100 at five years; available case and IPW included mean=337 and n=206 and LOCF used data from mean n=419 and n=302 with data recorded at at least one previous follow-up.

Mean prevalence rate in complete data=31% at one year and 32% at five years.

Abbreviations: se standard error, LOCF last observation carried forward, IPW Inverse probability weighting, MCAR missing completely at random, MAR missing at random, MNAR missing not at random.

Appendix D

Variation of parameter estimates between simulations

Table D.1: Standard deviation of estimates of prevalence of depression

		1 year				5 years			
		Sc I	Sc II	Sc III	Sc IV	Sc I	Sc II	Sc III	Sc IV
Categorical	Complete case	0.02592	0.02541	0.02518	0.02349	0.07944	0.07704	0.07919	0.0604
	Available case	0.01911	0.01899	0.01841	0.01948	0.02441	0.02343	0.02255	0.02326
	LOCF	0.01659	0.01746	0.0163	0.01674	0.01746	0.01698	0.01658	0.01893
	IPW	0.01913	0.01909	0.01857	0.01956	0.02555	0.02555	0.02387	0.02459
	Mode imputation	0.0111	0.01087	0.01022	0.01131	0.01308	0.01292	0.01241	0.01233
	Hotdeck imputation	0.02198	0.02167	0.02086	0.02261	0.02807	0.02788	0.02661	0.02577
	Regression imputation	0.02184	0.02261	0.02078	0.0204	0.02627	0.02582	0.02507	0.02259
	Multiple imputatiuon	0.01903	0.01933	0.0181	0.01947	0.02339	0.02257	0.02137	0.02210
Continuous	Median imputation	0.01095	0.01115	0.01098	0.01160	0.0131	0.01292	0.01253	0.01237
	Median imputation	0.01156	0.01114	0.01101	0.01157	0.0132	0.01283	0.01254	0.01252
	Hotdeck imputation	0.02647	0.02653	0.02521	0.02607	0.03932	0.0390	0.03764	0.03638
	Regression imputation	0.02316	0.02222	0.01996	0.02430	0.02158	0.02036	0.02109	0.01823
	Multiple imputation	0.01949	0.01948	0.01863	0.01956	0.02343	0.02287	0.02208	0.02264

Table displays the standard deviation of the point estimates of propotions with depression

Scenario I assumes outcome data are MCAR, Scenario II assumes MAR, Scenario III assumes MAR with missingness dependent on time of death and Scenario IV assumes MNAR data with missingness dependent on current disability level

Abbreviations: MCAR missing completely at random, MAR missing at random, MNAR missing not at random.

Table D.2: Standard deviation of estimates of prevalence of inactivity

		1 year				5 years			
		Sc I	Sc II	Sc III	Sc IV	Sc I	Sc II	Sc III	Sc IV
Categorical	Complete case	0.02207	0.0225	0.02239	0.03604	0.07241	0.05961	0.06594	0.09555
	Available case	0.01640	0.01636	0.01548	0.02659	0.02432	0.02313	0.02311	0.04126
	LOCF	0.01452	0.01391	0.01352	0.02114	0.01502	0.01506	0.01367	0.02302
	IPW	0.01556	0.01538	0.01466	0.02255	0.02371	0.0230	0.02173	0.03164
	Mode imputation	0.01593	0.01562	0.01673	0.05914	0.01951	0.01784	0.02156	0.09279
	Hotdeck imputation	0.02127	0.0212	0.02024	0.03062	0.03645	0.03509	0.0342	0.04741
	Regression imputation	0.02042	0.02003	0.01872	0.0253	0.03629	0.03622	0.03541	0.04308
	Multiple imputatiuon	0.01437	0.0141	0.01359	0.02073	0.01728	0.01704	0.01626	0.02378
Continuous	Median imputation	0.0392	0.0443	0.03217	0.02595	0.05507	0.06169	0.04466	0.03229
	Median imputation	0.05078	0.05693	0.04046	0.02974	0.08527	0.10241	0.09255	0.05537
	Hotdeck imputation	0.02433	0.02444	0.02484	0.03099	0.04274	0.04119	0.04025	0.05507
	Regression imputation	0.01738	0.01703	0.01634	0.02535	0.03053	0.0298	0.02908	0.03624
	Multiple imputation	0.01441	0.01477	0.01402	0.02064	0.01889	0.01807	0.01732	0.02685

Table displays the standard deviation of the point estimates of propotions who were inactive

Scenario I assumes outcome data are MCAR, Scenario II assumes MAR, Scenario III assumes MAR with missingness dependent on time of death and Scenario IV assumes MNAR data with missingness dependent on current disability level

Abbreviations:MCAR missing completely at random, MAR missing at random, MNAR missing not at random.

Table D.3: Standard deviation of estimates of prevalence of moderate-severe disability

		1 year				5 years			
		Sc I	Sc II	Sc III	Sc IV	Sc I	Sc II	Sc III	Sc IV
Categorical	Complete case	0.01964	0.0194	0.02021	0.02368	0.04826	0.04572	0.04656	0.0508
	Available case	0.01551	0.01491	0.01541	0.01925	0.02057	0.01989	0.01978	0.02913
	LOCF	0.01385	0.01323	0.01319	0.01738	0.01301	0.01289	0.01153	0.02297
	IPW	0.01477	0.01396	0.01435	0.01714	0.02001	0.01886	0.01858	0.02343
	Mode imputation	0.01786	0.01651	0.01648	0.01999	0.01468	0.0138	0.01435	0.01782
	Hotdeck imputation	0.01953	0.01846	0.01903	0.02109	0.02383	0.02188	0.02171	0.02653
	Regression imputation	0.01694	0.01548	0.01625	0.01864	0.02001	0.01852	0.01882	0.02757
	Multiple imputatiuon	0.01370	0.01332	0.01293	0.01537	0.01471	0.01425	0.0139	0.0170
Continuous	Median imputation	0.04196	0.0398	0.03853	0.04403	0.04822	0.04594	0.04437	0.04347
	Median imputation	0.01844	0.01726	0.01755	0.02322	0.01581	0.01444	0.01516	0.02178
	Hotdeck imputation	0.02407	0.0236	0.02307	0.02462	0.03239	0.02993	0.03058	0.03774
	Regression imputation	0.01747	0.01672	0.01671	0.01801	0.02784	0.02505	0.02508	0.02366
	Multiple imputation	0.01410	0.01351	0.01313	0.01778	0.01582	0.01547	0.01476	0.04696

Table displays the standard deviation of the point estimates of propotions with moderate-severe disability

Scenario I assumes outcome data are MCAR, Scenario II assumes MAR, Scenario III assumes MAR with missingness dependent on time of death and Scenario IV assumes MNAR data with missingness dependent on current disability level

Abbreviations: MCAR missing completely at random, MAR missing at random, MNAR missing not at random.

Table D.4: Standard deviation of estimates of prevalence of anxiety

		1 year				5 years			
		Sc I	Sc II	Sc III	Sc IV	Sc I	Sc II	Sc III	Sc IV
Categorical	Complete case	0.02646	0.02695	0.02672	0.02332	0.0836	0.07839	0.07972	0.0606
	Available case	0.01975	0.0204	0.01848	0.01891	0.02412	0.02415	0.0238	0.02131
	LOCF	0.01775	0.0179	0.01588	0.01695	0.0176	0.01749	0.01711	0.01816
	IPW	0.01993	0.02061	0.01886	0.01942	0.02482	0.02549	0.02454	0.02181
	Mode imputation	0.01201	0.01281	0.01441	0.01266	0.01402	0.01469	0.01575	0.01269
	Hotdeck imputation	0.02264	0.02507	0.02225	0.02394	0.03213	0.03253	0.03209	0.0271
	Regression imputation	0.02973	0.02986	0.02863	0.03081	0.03852	0.04016	0.0389	0.03512
	Multiple imputatiuon	0.02012	0.02025	0.01814	0.01894	0.02235	0.02235	0.02132	0.01973
Continuous	Median imputation	0.01075	0.01058	0.01006	0.01121	0.01353	0.01356	0.01331	0.01204
	Median imputation	0.01218	0.01281	0.01441	0.01266	0.03431	0.03594	0.03385	0.01875
	Hotdeck imputation	0.03194	0.03351	0.03151	0.03215	0.03962	0.04036	0.03987	0.03764
	Regression imputation	0.02327	0.02436	0.02343	0.02389	0.03016	0.03036	0.03046	0.02729
	Multiple imputation	0.02023	0.02032	0.01873	0.01965	0.02263	0.02272	0.02156	0.02039

Table displays the standard deviation of the point estimates of propotions with anxiety

Scenario I assumes outcome data are MCAR, Scenario II assumes MAR, Scenario III assumes MAR with missingness dependent on time of death and Scenario IV assumes MNAR data with missingness dependent on current disability level

Abbreviations: MCAR missing completely at random, MAR missing at random, MNAR missing not at random.

Appendix E

Effect of missing data on predictors of anxiety after stroke

Table E.1: Completeness of HADS Anxiety measurements and prevalence of anxiety in SLSR participants 1995-2007

	3 months	1 year	2 years	3 years	4 years	5 years
Total alive	2615	2320	2123	1917	1782	1654
Completed anxiety measurement, n(%)	950(36.3)	1070(46.1)	840(39.6)	1134(59.2)	979(55.4)	808(49.9)
Not anxious	639(67.3)	748(69.9)	572(68.1)	785(69.2)	667(68.1)	555(68.7)
Anxious	311(32.7)	322(30.1)	268(31.9)	349(30.8)	312(31.9)	253(31.3)
Reason for missing measurement, n(%)						
Lost to follow-up	783(46.9)	595(47.6)	534(41.6)	571(72.9)	578(73.4)	627(77.2)
HADS not on form	609(36.5)	371(29.7)	521(40.6)	0	0	0
HADS not done for other reason	276(16.6)	284(22.7)	228(17.8)	212(27.1)	210(26.7)	185(22.8)

Abbreviations: HADS Hospital anxiety and depression scale.

Table E.2: Relationship between time after stroke and anxiety

	beta	standard error	t-value	p-value
time	0.015	0.024	0.63	0.527
time	-0.168	0.090	-0.19	0.852
$time^2$	0.006	0.017	0.37	0.710
time	0.014	0.232	0.06	0.954
$time^2$	-0.008	0.106	-0.08	0.937
$time^3$	0.002	0.013	0.14	0.888

Table shows parameter estimates from logistic GLMMs with random intercept for the relationship between anxiety and time since stroke. Models were adjusted for age, sex, ethnicity, stroke subtype, Glasgow coma score and disability 7-10 days after stroke.

Table E.3: Unadjusted logistic GEE approach models exploring the associations between baseline characteristics and post stroke anxiety

	GEE				WGEE (mortal)				WGEE (immortal)				MIGEE			
	beta	se	z	p-value	beta	se	z	p-value	beta	se	z	p-value	beta	se	t	p-value
Age	-0.011	0.003	-3.46	0.001	-0.018	0.002	-8.91	<0.001	-0.017	0.002	-8.22	<0.001	-0.012	0.005	-2.49	0.016
Sex																
Male	ref				ref				ref				ref			
Female	0.336	0.047	3.97	<0.001	0.166	0.052	3.19	0.001	0.153	0.053	2.88	0.004	0.214	0.134	1.60	0.124
Ethnicity																
White	ref				ref				ref				ref			
Black	-0.184	0.101	-1.81	0.07	-0.207	0.065	-3.21	0.001	-0.219	0.063	-3.48	0.001	-0.099	0.15	-0.66	0.509
Other	0.02	0.167	0.12	0.906	0.011	0.093	0.13	0.475	0.115	0.133	0.86	0.388	0.091	0.224	0.41	0.682
Subtype																
Infarct	ref				ref				ref				ref			
PICH	-0.163	0.147	-1.11	0.268	-0.112	0.085	1.32	0.186	-0.095	0.088	1.09	0.275	-0.066	0.212	-0.31	0.760
SAH	0.246	0.185	1.33	0.184	0.156	0.102	-1.53	0.127	0.164	0.111	1.48	0.139	0.147	0.333	0.44	0.660
Undefined	0.183	0.189	0.97	0.333	0.028	0.144	-0.19	0.847	0.039	0.146	0.27	0.788	0.190	0.287	0.66	0.511
GCS	-0.01	0.015	-0.66	0.506	-0.019	0.01	-1.80	0.072	-0.032	0.011	-2.89	0.004	-0.026	0.021	-1.22	0.23
7-10d Disability																
Severe	ref				ref				ref				ref			
Moderate	0.086	0.132	0.65	0.518	0.089	0.084	1.06	0.288	0.049	0.084	0.59	0.556	0.058	0.152	0.38	0.705
Mild	-0.054	0.120	-0.46	0.649	-0.046	0.08	-0.57	0.568	0.005	0.077	0.06	0.952	-0.022	0.152	-0.14	0.862
Indep	-0.194	0.105	-1.84	0.066	-0.177	0.064	-2.75	0.006	-0.226	0.064	3.51	<0.001	-0.335	0.146	-2.43	0.018

All models were adjusted for time since stroke as a linear covariate.

Abbreviations: GEE generalised estimation equations, WGEE weighted generalised estimating equations, MIGEE generalised estimating equations with multiple imputation, se standard error, GCS Glasgow coma score, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

Table E.4: Unadjusted likelihood based logistic models exploring the association between baseline characteristics and post stroke anxiety

	GLMM				Shared parameter model							Pattern mixture			
									Shared random effect						
	beta	se	t	p-value	beta	se	t	p-value	Var	se	p-value	beta	se	t	p-value
Age	-0.013	0.004	-3.41	<0.001	-0.02	0.005	-4.04	<0.001	1.167	0.090	<0.001	-0.012	0.007	-1.71	0.087
Sex									1.159	0.090	<0.001				
Male					ref							ref			
Female	0.410	0.104	3.93	<0.001	0.569	0.137	4.17	<0.001			<0.001	0.354	0.14	2.53	0.012
Ethnicity									1.171	0.092	<0.001				
White	ref				ref							ref			
Black	-0.219	0.127	-1.73	0.084	-0.204	0.162	-1.26	0.207				-0.209	0.17	-1.23	0.219
Other	0.085	0.194	0.44	0.66	-0.061	0.289	-0.21	0.834				0.047	0.303	0.16	0.877
Subtype									1.175	0.090	<0.001				
Infarct	ref				ref							ref			
PICH	-0.201	0.177	-1.13	0.257	-0.103	0.225	-0.46	0.648				-0.103	0.253	-0.41	0.684
SAH	0.302	0.234	1.29	0.196	0.451	0.32	1.41	0.159				0.37	0.337	1.10	0.272
Undefined	0.227	0.229	0.99	0.322	0.371	0.302	1.23	0.219				0.364	0.345	1.05	0.292
GCS	-0.045	0.02	-2.27	0.024	-0.019	0.027	-0.72	0.469	1.177	0.092	<0.001	-0.01	0.03	-0.34	0.734
7-10d Disability									1.167	0.090	<0.001				
Severe	ref				ref							ref			
Moderate	0.094	0.161	0.59	0.558	0.285	0.215	1.33	0.185				0.244	0.222	1.10	0.271
Mild	-0.071	0.148	-0.48	0.633	-0.025	0.196	-0.13	0.899				-0.017	0.166	-0.10	0.92
Independent	-0.236	0.13	-1.82	0.069	-0.134	0.169	-0.79	0.429				-0.153	0.142	-1.08	0.282

All models were adjusted for time since stroke as a linear covariate.

Abbreviations: GLMM generalised linear mixed model, se standard error, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

Table E.5: Unadjusted logistic pattern mixture models for the association between baseline characteristics and post stroke anxiety (part 1 of 2)

	Main effect				max time = 3m				max time = 1yr			
	beta	se	t	p-value	beta	se	t	p-value	beta	se	t	p-value
Age	-0.010	0.006	-1.82	0.069	0.013	0.026	0.48	0.632	0.005	0.017	0.28	0.783
Sex												
Male					ref				ref			
Female	0.392	0.146	2.7	0.007	-0.860	0.723	-1.19	0.234	0.371	0.488	0.76	0.447
Ethnicity												
White	ref				ref				ref			
Black	-0.170	0.152	-1.12	0.263	0.013	0.491	0.03	0.980	-0.12	0.351	-0.34	0.732
Other	0.091	0.254	0.36	0.720	0.002	0.641	0.000	0.998	-0.007	0.414	-0.02	0.987
Subtype												
Infarct	ref				ref				ref			
PICH	-0.081	0.176	-0.46	0.644	-0.002	0.568	-0.00	0.997	-0.011	0.521	-0.02	0.984
SAH	0.412	0.301	1.37	0.171	0.078	0.895	0.09	0.931	-0.475	0.785	-0.60	0.545
Undefined	0.303	0.287	1.06	0.291	0.021	0.745	0.03	0.977	0.421	0.635	0.66	0.507
GCS	0.015	0.029	0.54	0.592	0.024	0.174	0.14	0.890	-0.034	0.104	-0.33	0.742
7-10d Disability												
Severe	ref				ref				ref			
Moderate	0.201	0.204	0.99	0.323	0.321	0.569	0.56	0.572	-0.003	0.561	-0.000	0.996
Mild	0.001	0.12	0.01	0.993	-0.075	0.325	-0.23	0.817	-0.012	0.313	-0.04	0.969
Independent	-0.156	0.158	-0.99	0.323	0.022	0.356	0.06	0.952	0.000	0.345	0.012	0.999

Pattern mixture model showing the association between covariates and anxiety in those with complete data (the main effect) and the additional effect the parameters have in those who dropped out broken down by the maximum follow-up time of those without five years of follow-ups. All models were adjusted for age as a linear covariate.

Abbreviations: se standard error, GCS Glasgow coma score, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

Table 7.7 Unadjusted logistic pattern mixture models for the association between baseline characteristics and post stroke anxiety (part 2 of 2)

	max time = 2yr				max time = 3yr				max time = 4yr			
	beta	se	t	p-value	beta	se	t	p-value	beta	se	t	p-value
Age	-0.003	0.014	-0.23	0.817	-0.005	0.011	-0.41	0.680	-0.018	0.011	-1.67	0.096
Sex												
Male	ref				ref				ref			
Female	0.071	0.376	0.19	0.851	-0.024	0.331	-0.07	0.942	0.046	0.272	0.17	0.867
Ethnicity												
White	ref				ref				ref			
Black	0.11	0.35	0.31	0.753	0.001	0.321	0.00	0.998	-0.214	0.274	-0.78	0.435
Other	0.109	0.401	0.27	0.786	-0.215	0.329	-0.65	0.514	-0.152	0.325	-0.47	0.639
Subtype												
Infarct	ref				ref				ref			
PICH	-0.011	0.4	-0.03	0.977	-0.012	0.351	-0.04	0.972	-0.102	0.215	-0.48	0.634
SAH	0.100	0.713	0.14	0.888	0.063	0.665	0.09	0.925	-0.055	0.541	-0.10	0.919
Undefined	-0.022	0.621	-0.04	0.972	0.025	0.597	0.04	0.966	0.022	0.524	0.04	0.966
GCS	-0.045	0.02	-2.27	0.024	-0.019	0.027	-0.72	0.469	-0.01	0.03	-0.34	0.734
7-10d Disability												
Severe	ref				ref				ref			
Moderate	-0.125	0.501	-0.25	0.803	0.021	0.423	0.05	0.96	0.100	0.345	0.29	0.772
Mild	-0.102	0.275	-0.37	0.710	0.02	0.254	0.08	0.936	0.000	0.185	0.00	1.00
Indep	-0.056	0.281	-0.20	0.841	0.022	0.265	0.08	0.934	0.022	0.206	0.10	0.917

Pattern mixture model showing the association between covariates and anxiety in those with complete data (the main effect) and the additional effect the parameters have in those who dropped out broken down by the maximum follow-up time of those without five years of follow-ups. All models were adjusted for age as a linear covariate.

Abbreviations: se standard error, GCS Glasgow coma score, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

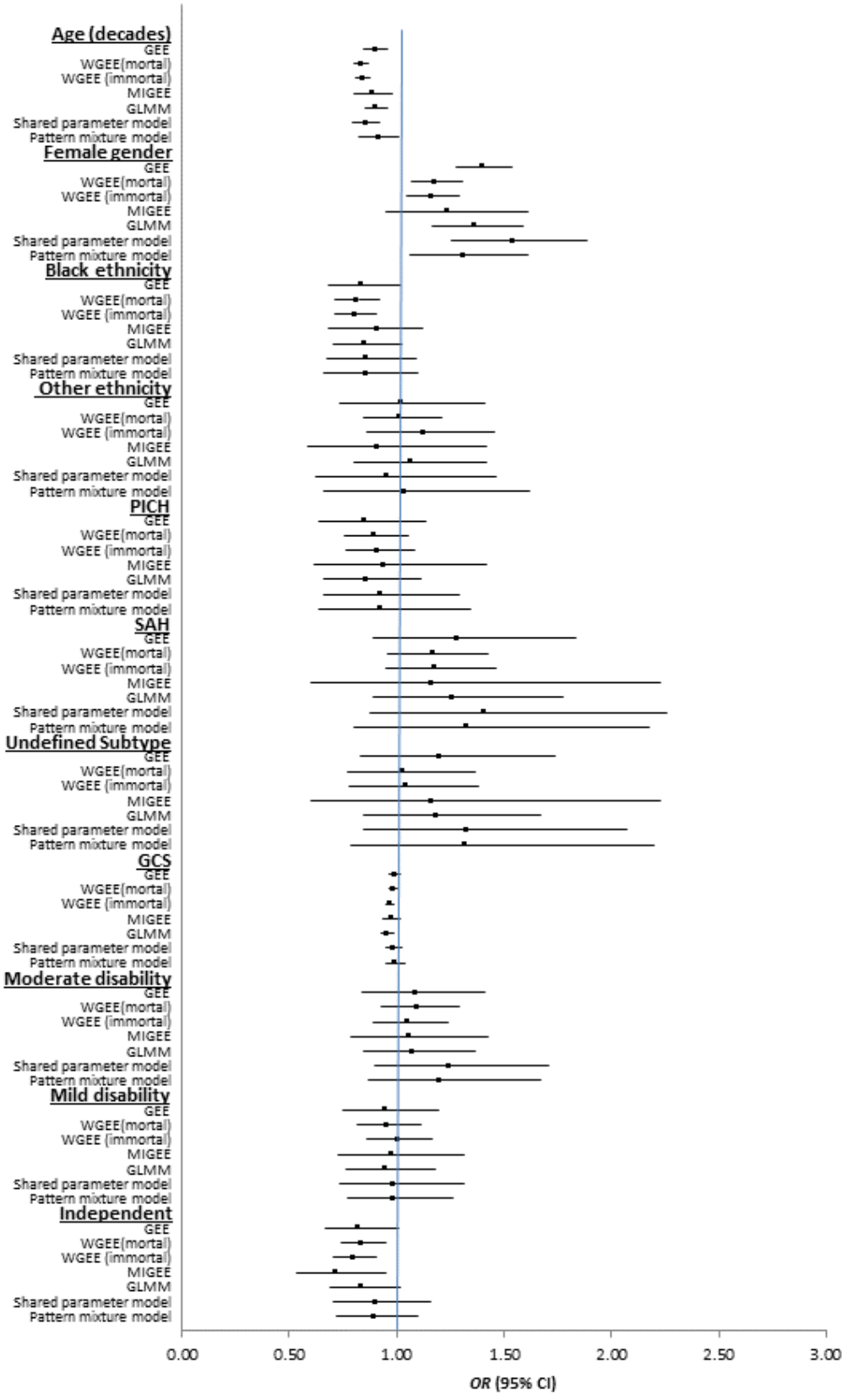


Figure E.1: Unadjusted estimates of the marginal effect of baseline characteristics on the odds of post stroke anxiety

Table E.6: Adjusted logistic GEE approach models exploring the associations between baseline characteristics and post stroke anxiety

	GEE				WGEE (mortal)				WGEE (immortal)				MIGEE			
	beta	se	z	p-value	beta	se	z	p-value	beta	se	z	p-value	beta	se	t	p-value
Age	-0.02	0.003	-4.49	<0.001	-0.028	0.002	-11.9	<0.00	-0.026	0.002	-11.39	<0.001	-0.017	0.005	-3.24	0.002
Sex																
Male	ref				ref				ref				ref			
Female	0.375	0.086	4.32	<0.001	0.247	0.049	5.09	<0.00	0.23	0.051	4.55	<0.001	0.263	0.136	1.94	0.065
Ethnicity																
White	ref				ref				ref				ref			
Black	-0.32	0.108	-2.96	0.003	-0.426	0.069	-6.13	0	-0.429	0.068	-6.29	<0.00	-0.38	0.151	-2.52	0.434
Other	-0.056	0.17	-0.33	0.741	-0.272	0.138	-1.97	0.049	-0.296	0.14	-2.12	0.034	-0.185	0.235	-0.79	0.652
Subtype																
Infarct	ref				ref				ref				ref			
PICH	-0.256	0.157	-1.64	0.102	-0.315	0.093	-3.4	0.001	-0.311	0.097	-3.2	0.001	-0.113	0.211	-0.54	0.595
SAH	-0.089	0.206	-0.43	0.665	-0.397	0.104	-3.8	<0.00	-0.384	0.113	-3.45	0.001	-0.344	0.396	-0.87	0.385
Undefined	0.187	0.185	1.01	0.313	0.038	0.135	0.28	0.777	0.047	0.138	0.34	0.731	-0.116	0.295	-0.39	0.695
GCS	0.004	0.017	0.25	0.803	-0.001	0.012	-0.06	0.950	0.016	0.013	1.27	0.203	0.015	0.029	-0.52	0.611
7-10d Disability																
Severe	ref				ref				ref				ref			
Moderate	-0.256	0.157	0.79	0.102	0.174	0.087	2.00	0.045	0.158	0.086	1.84	0.066	0.127	0.174	0.07	0.485
Mild	-0.089	0.206	-0.39	0.665	0.042	0.082	0.51	0.611	0.022	0.081	0.27	0.786	0.015	0.179	0.09	0.933
Indep	-0.187	0.186	-2.06	0.313	-0.302	0.073	-4.14	<0.001	-0.32	0.073	-4.38	<0.001	-0.313	0.187	-1.67	0.106

All models were adjusted for time since stroke as a linear covariate.

Abbreviations: GEE generalised estimation equations, WGEE weighted generalised estimating equations, MIGEE generalised estimating equations with multiple imputation, se standard error, GCS Glasgow coma score, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

Table E.7: Adjusted likelihood based logistic models exploring the association between baseline characteristics and post stroke anxiety

	GLMM				Shared parameter model						
									Shared random effect		
	beta	se	t	p-value	beta	se	t	p-value	Var	se	p-value
Age	-0.019	0.004	-4.6	<0.001	-0.026	0.006	-4.33	<0.001	1.118	0.089	<0.001
Sex											
Male	ref				ref						
Female	0.455	0.105	4.33	<0.001	0.642	0.138	4.65	<0.001			
Ethnicity											
White	ref				ref						
Black	-0.368	0.129	-2.86	0.004	-0.428	0.165	-2.59	0.010			
Other	0.002	0.193	0.01	0.990	-0.175	0.288	-0.61	0.542			
Subtype											
Infarct	ref				ref						
PICH	-0.605	0.177	-1.72	0.086	-0.238	0.227	-1.05	0.294			
SAH	-0.102	0.242	-0.42	0.675	-0.078	0.333	-0.23	0.818			
Undefined	0.218	0.227	0.96	0.336	0.359	0.299	1.20	0.230			
GCS	0.004	0.022	0.19	0.849	-0.009	0.03	-0.30	0.764			
7-10d Disability											
Severe	ref				ref						
Moderate	0.121	0.167	0.73	0.468	0.379	0.223	1.7	0.089			
Mild	-0.070	0.155	-0.45	0.650	0.085	0.205	0.41	0.682			
Independent	-0.298	0.142	-2.10	0.036	-0.147	0.186	-0.79	0.430			

All models were adjusted for time since stroke as a linear covariate.

Abbreviations: GLMM generalised linear mixed model, se standard error, PICH primary intracerebral haemorrhage, SAH subarachnoid haemorrhage.

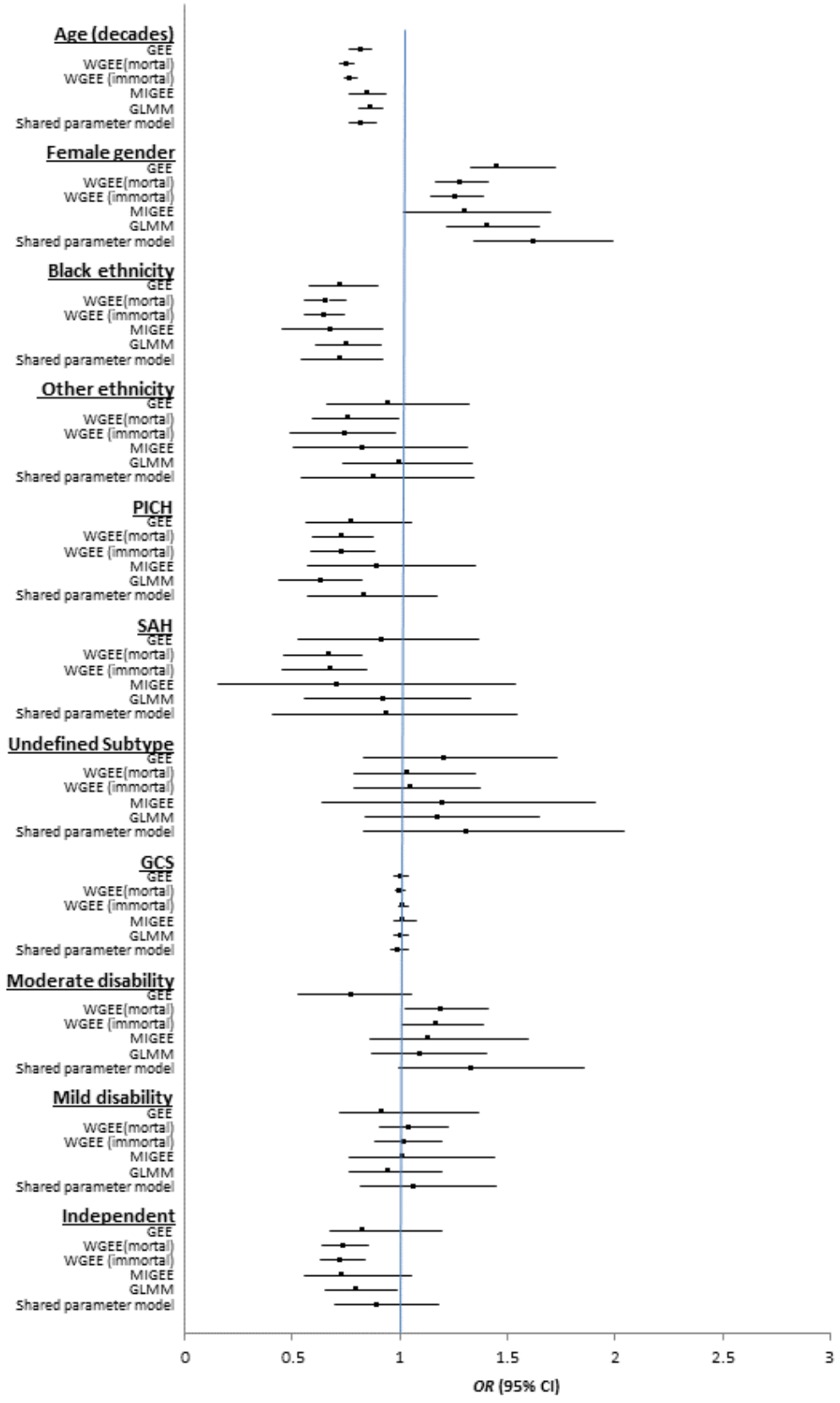


Figure E.2: Adjusted estimates of the marginal effect of baseline characteristics on the odds of post stroke anxiety